AN INTRODUCTION TO MISSING DATA ANALYSES

Craig K. Enders

University of California Los Angeles

Department of Psychology

WWW.APPLIEDMISSINGDATA.COM/BLIMP-PAPERS

APPLIED MISSING DATA

home analysis examples blimp blimp papers videos centerstat workshop quantitude podcast

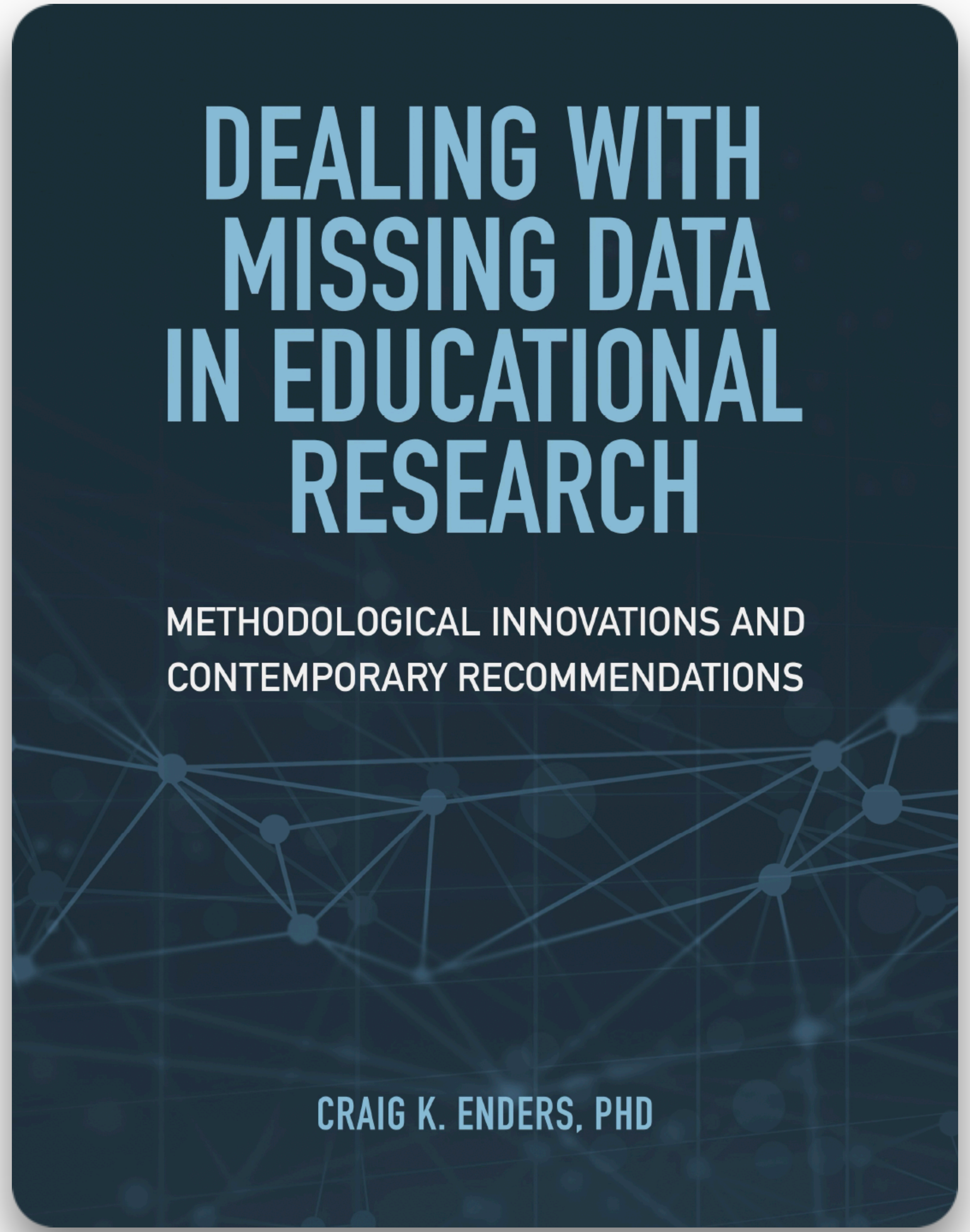
Workshops and Training

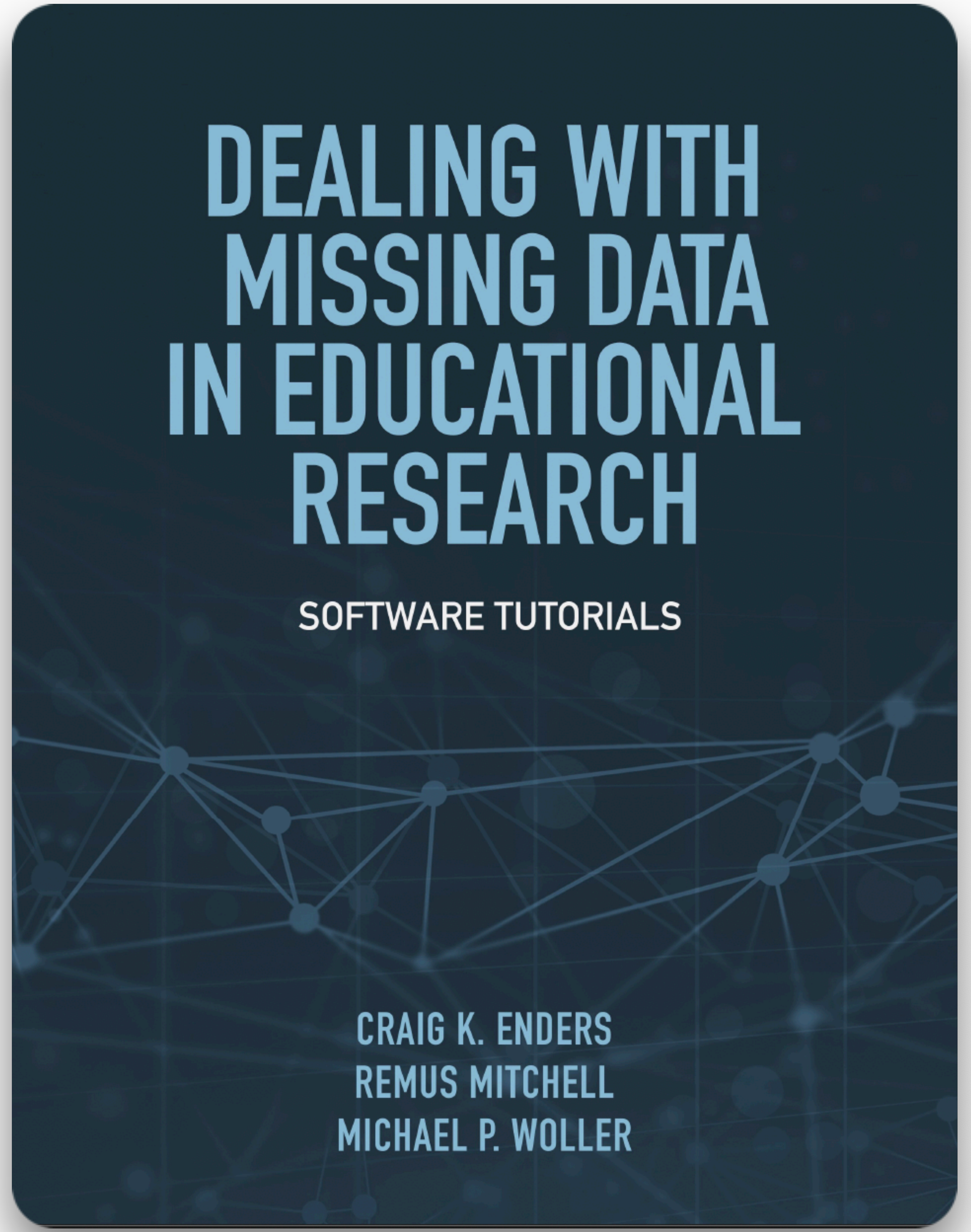
Enders, C. K. (2025, March). *An introduction to missing data analyses* [Webinar]. Johannes Kepler University Linz, Austria.

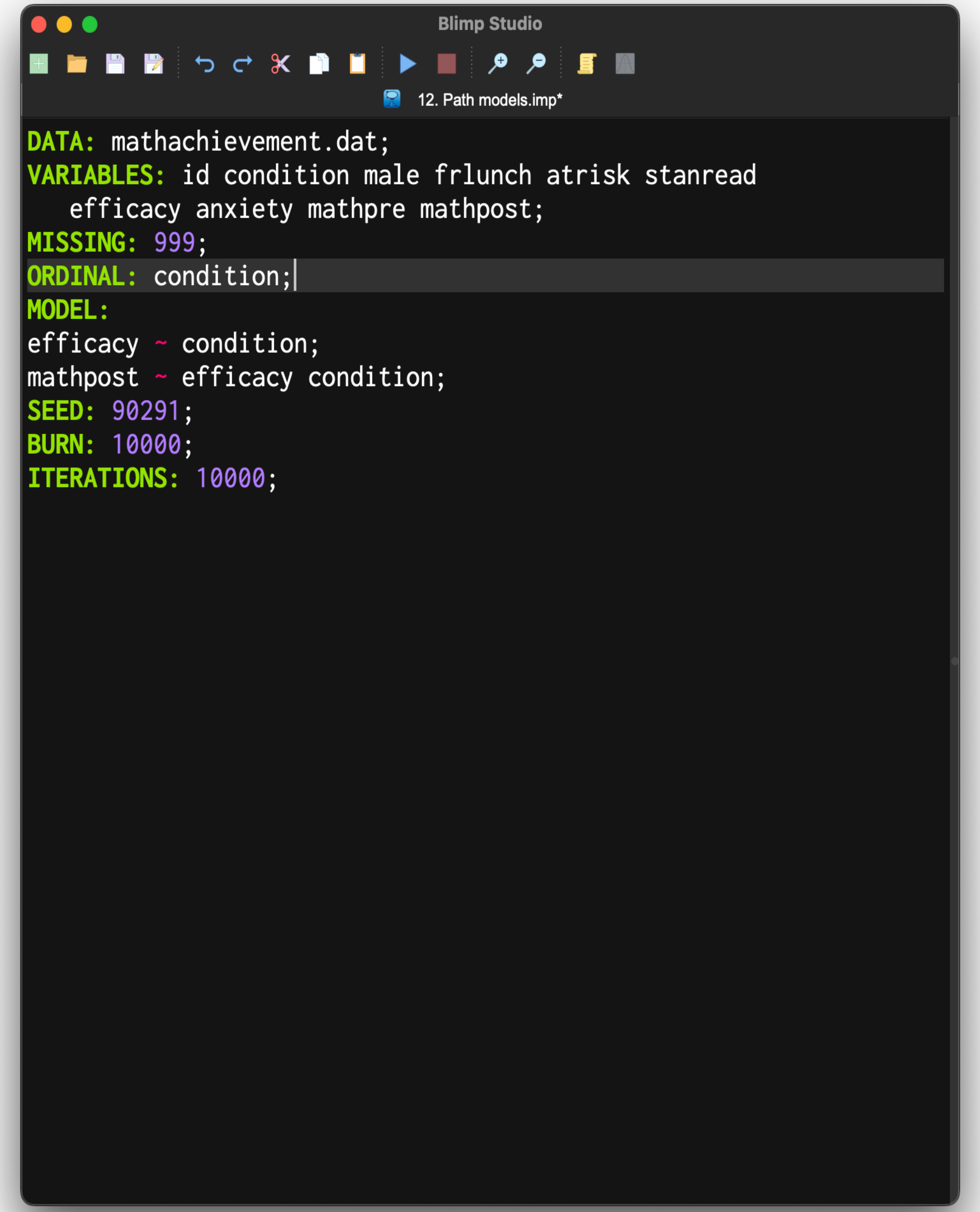
DOWNLOAD WEBINAR MATERIALS

RESOURCES

IES Institute of Education Sciences







OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Bayesian MCMC Estimation
5. Multiple Imputation
6. Missing Data Software
7. Analysis Example

OUTLINE

MODERN MISSING DATA METHODS

Maximum likelihood

Bayesian estimation

Multiple imputation

the Big Three

KEY ADVANTAGES OF BIG THREE

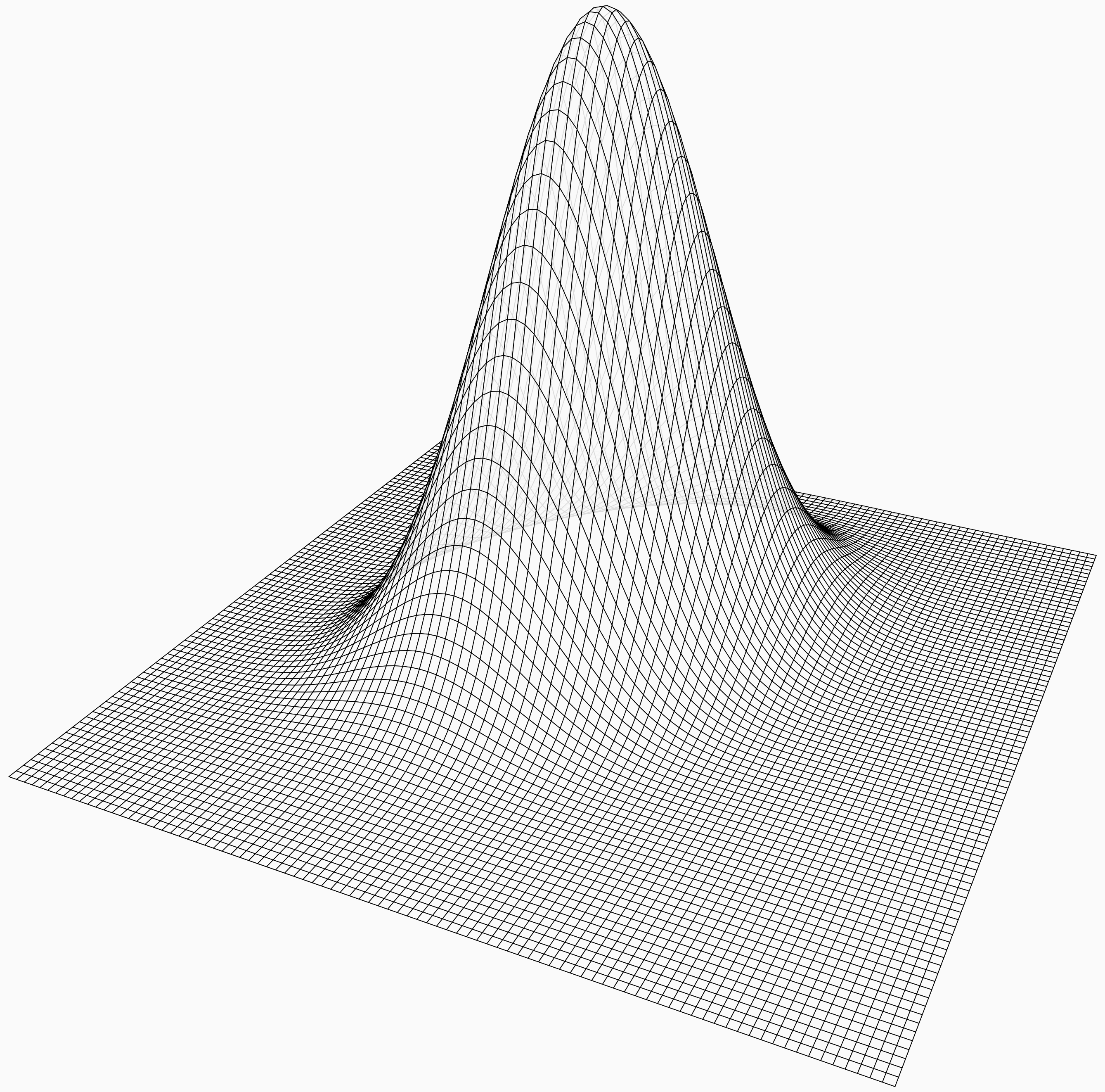
- Achieve unbiasedness with more a realistic assumption about the missing data process
- Allow for alternate assumptions about nonresponse process
- Maximize power
- Use all available data, no wasted resources

CHOOSING A MISSING DATA METHOD

- All things being equal—same data, same variables, same assumptions—the Big Three rarely produce different results
- Missing data analyses require distributional assumptions
- How we represent those distributions—multivariate versus factored specifications—is what matters

MODELING FRAMEWORKS

Multivariate modeling

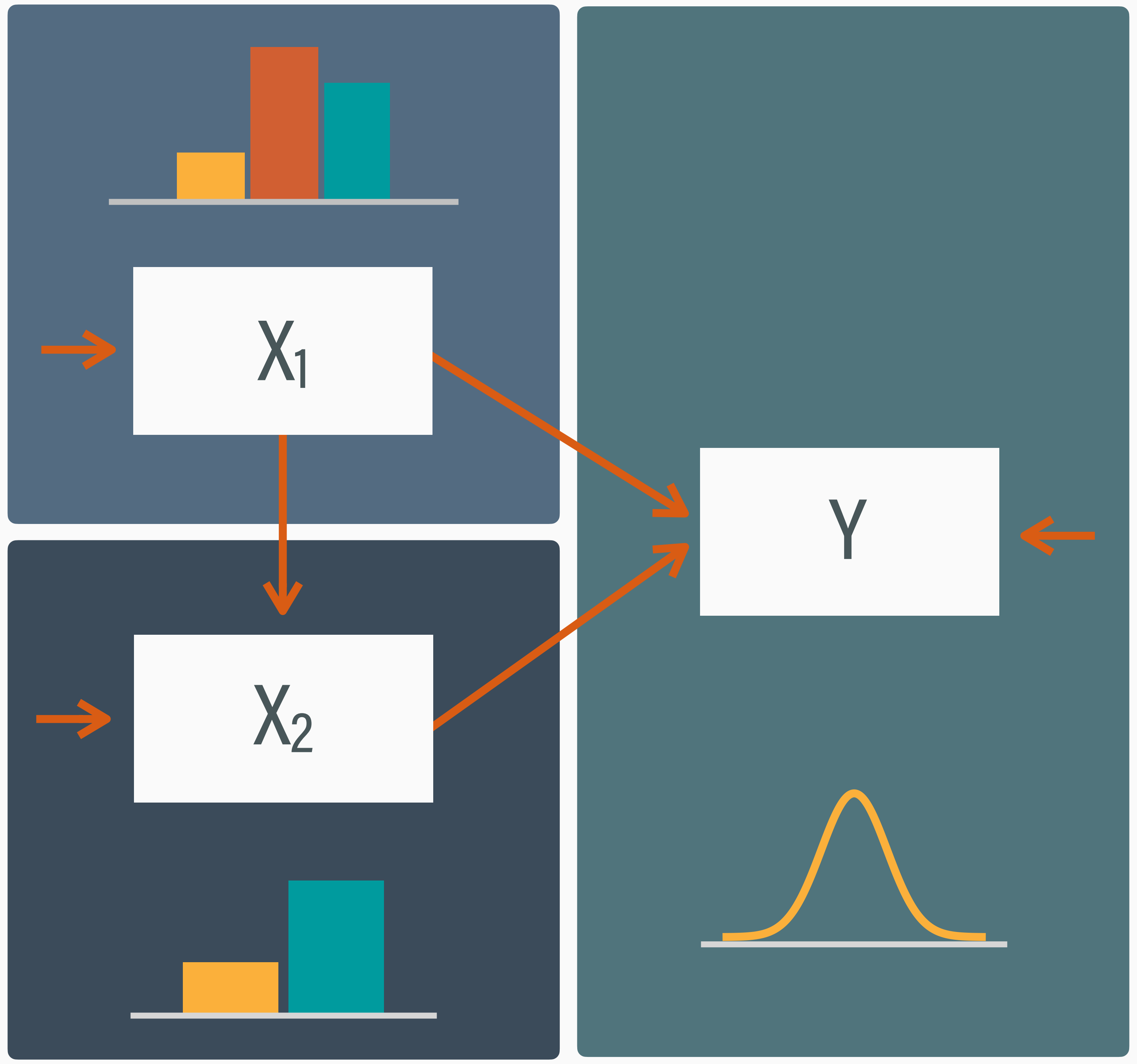


- Classic approaches often assume multivariate normality
- Most applications of maximum likelihood and multiple imputation

MODELING FRAMEWORKS

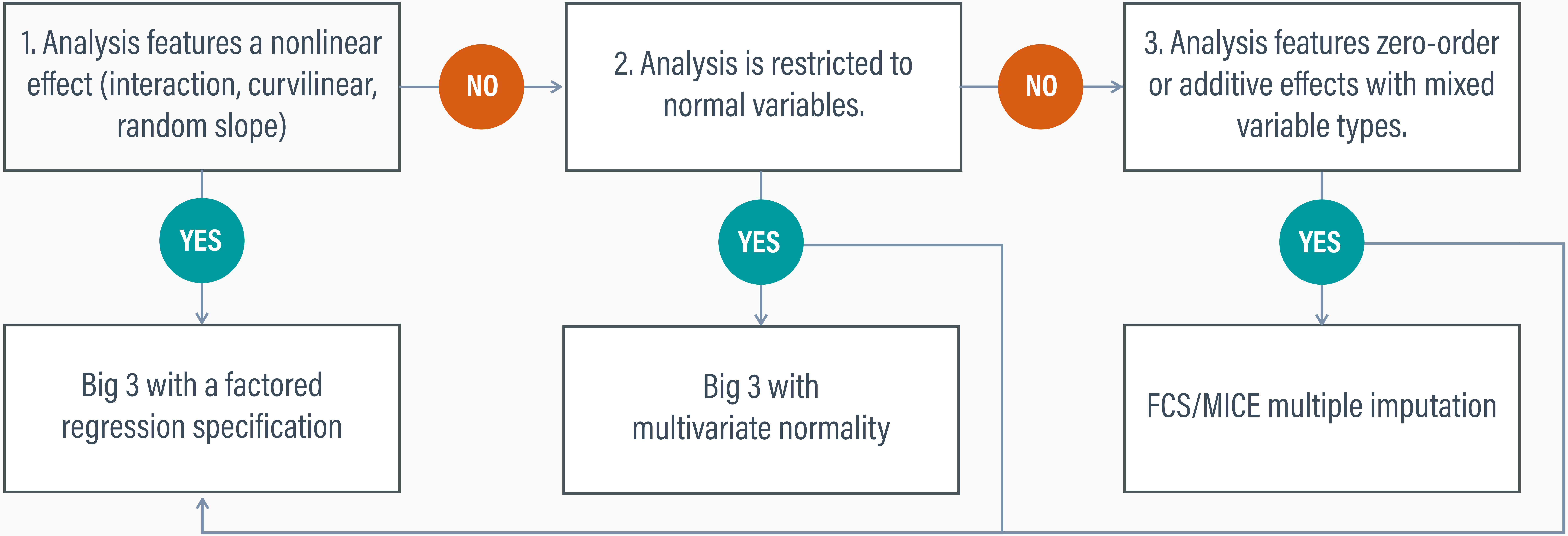
Multivariate modeling

Factored regression specification



- Factored regression invokes a unique model and distribution for each variable
- Each model can include terms that are at odds with multivariate normality (e.g., interactions, random slopes)

MISSING DATA DECISION TREE



OUTLINE

HOW MUCH MISSING DATA IS TOO MUCH?

- The Big Three can tolerate substantial amounts of missing data
- The Big Three are increasingly better than ad hoc methods (e.g., deleting incomplete cases) as missingness increases
- The amount of missing data is less important than why the data are missing (the missingness process or mechanism)

MISSING DATA MECHANISMS

- Missing data mechanisms (processes) describe different ways in which the data relate to nonresponse
- Missingness may be completely random or systematically related to different parts of the data
- Mechanisms function as statistical assumptions

PARTITIONING THE DATA

Complete = Observed + Missing | Indicators

Y_1	Y_2	Y_3		Y_1	Y_2	Y_3		Y_1	Y_2	Y_3	M_1	M_2	M_3
4	4	3		4	4	3					0	0	0
3	**3**	5		3	**NA**	5			**3**		0	**1**	0
7	1	6		7	1	6					0	0	0
2	1	6		**NA**	1	6		**2**			**1**	0	0
5	9	3	=	5	9	3	+				0	0	0
3	**2**	**2**		3	**NA**	**NA**			**2**	**2**	0	**1**	**1**
1	6	7		1	6	7					0	0	0
9	4	9		9	4	9					0	0	0
2	**5**	6		2	**NA**	6			**5**		0	**1**	0

MISSING COMPLETELY AT RANDOM

- The probability of missing values is completely unrelated to the data
- MCAR is purely random missingness
- We don't care about this process or testing for it (e.g., Little's MCAR test)

Missingness

Indicators

M_1	M_2	M_3
0	0	0
0	1	0
0	0	0
1	0	0
0	0	0
0	1	1
0	0	0
0	0	0
0	1	0

Predictors of nonresponse

Observed

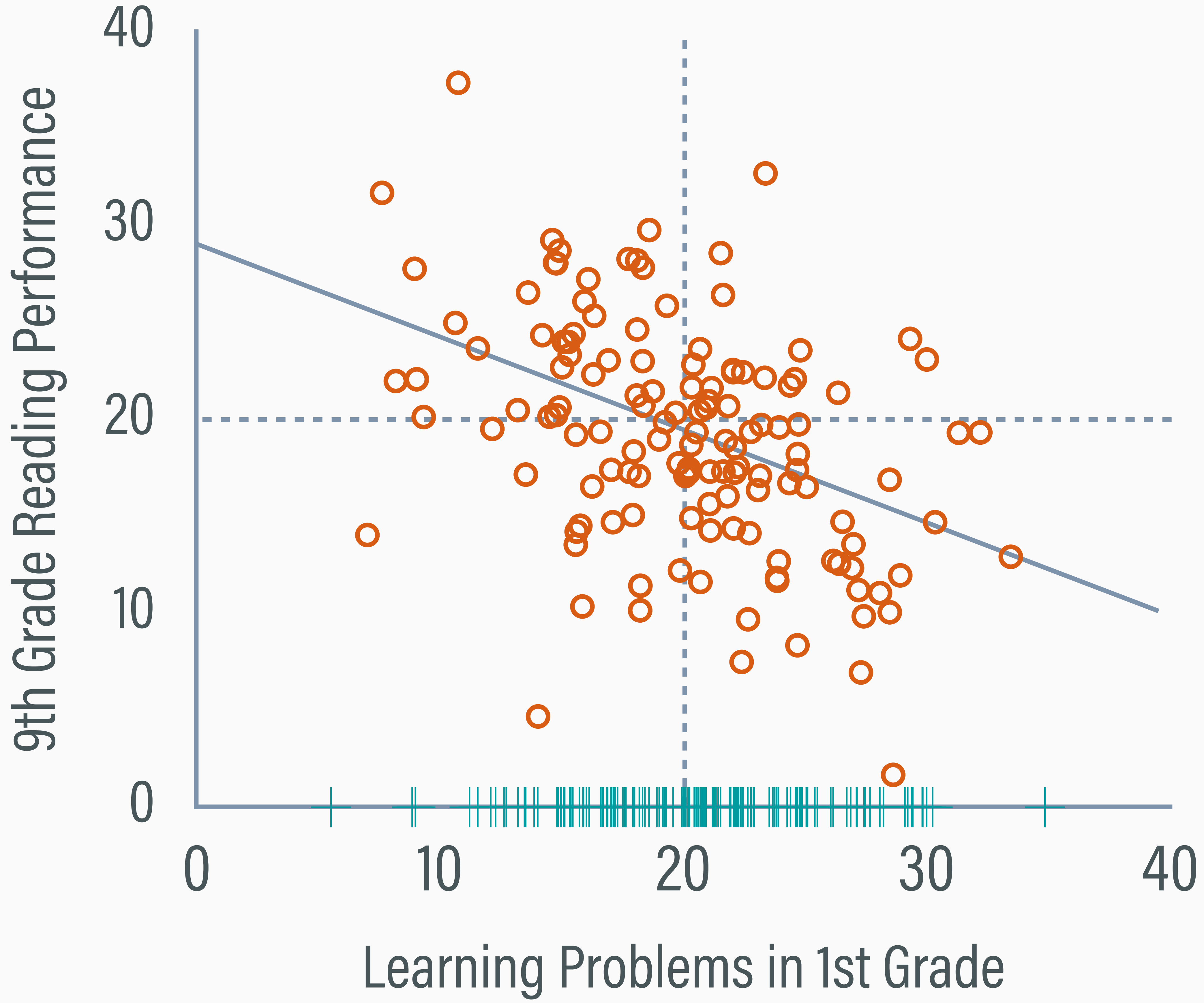
Y_1	Y_2	Y_3
4	4	3
3	NA	5
7	1	6
NA	[illegible]	6
5	9	3
3	NA	NA
1	6	7
9	4	9
2	NA	6

Missing

Y_1	Y_2	Y_3
	3	
2	[illegible]	[illegible]
	2	2
	5	

RESEARCH SCENARIO

- Study investigating association between learning problems in 1st grade and reading performance in 9th grade
- Learning problems ratings are complete and reading scores are incomplete



MCAR EXAMPLE

- Missingness is unrelated to the observed learning problems measure and unrelated to the unseen reading scores
- Planned missing data design where 9th grade reading scores are collected from a random subset of the original sample in order to reduce data collection costs
- Unplanned missingness is unrelated to the data (e.g., scheduling conflicts, administrative or logistical errors, family relocation)

(CONDITIONALLY) MISSING AT RANDOM

- Systematic missingness related to the observed scores
- The probability of missing values is unrelated to the unseen (latent) data
- Most Big Three applications assume CMAR

Missingness			Predictors of nonresponse					
Indicators			Observed			Missing		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	1	0	3	NA	5		3	
0	0	0	7	1	6			
1	0	0	NA	1	6	2		
0	0	0	5	9	3			
0	1	1	3	NA	NA		2	2
0	0	0	1	6	7			
0	0	0	9	4	9			
0	1	0	2	NA	6		5	

CONDITIONALLY MAR EXAMPLE

- Missingness is related to the observed learning problems measure but unrelated to the unseen reading scores
- Students with high levels of learnings problems are more likely to have missing data due to increased dropout risk, disciplinary actions, or family or situational instability
- The Big Three assume a CMAR process by default

MISSING NOT AT RANDOM

- Systematic missingness
- The probability of missing values is related to the unseen (latent) data
- The Big Three also allow MNAR processes (selection and pattern mixture models)

Missingness			Predictors of nonresponse					
Indicators			Observed			Missing		
M_1	M_2	M_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
0	0	0	4	4	3			
0	**1**	0	3	**NA**	5		**3**	
0	0	0	7	1	6			
1	0	0	**NA**	1	6	**2**		
0	0	0	5	9	3			
0	**1**	**1**	3	**NA**	**NA**		**2**	**2**
0	0	0	1	6	7			
0	0	0	9	4	9			
0	**1**	0	2	**NA**	6		**5**	

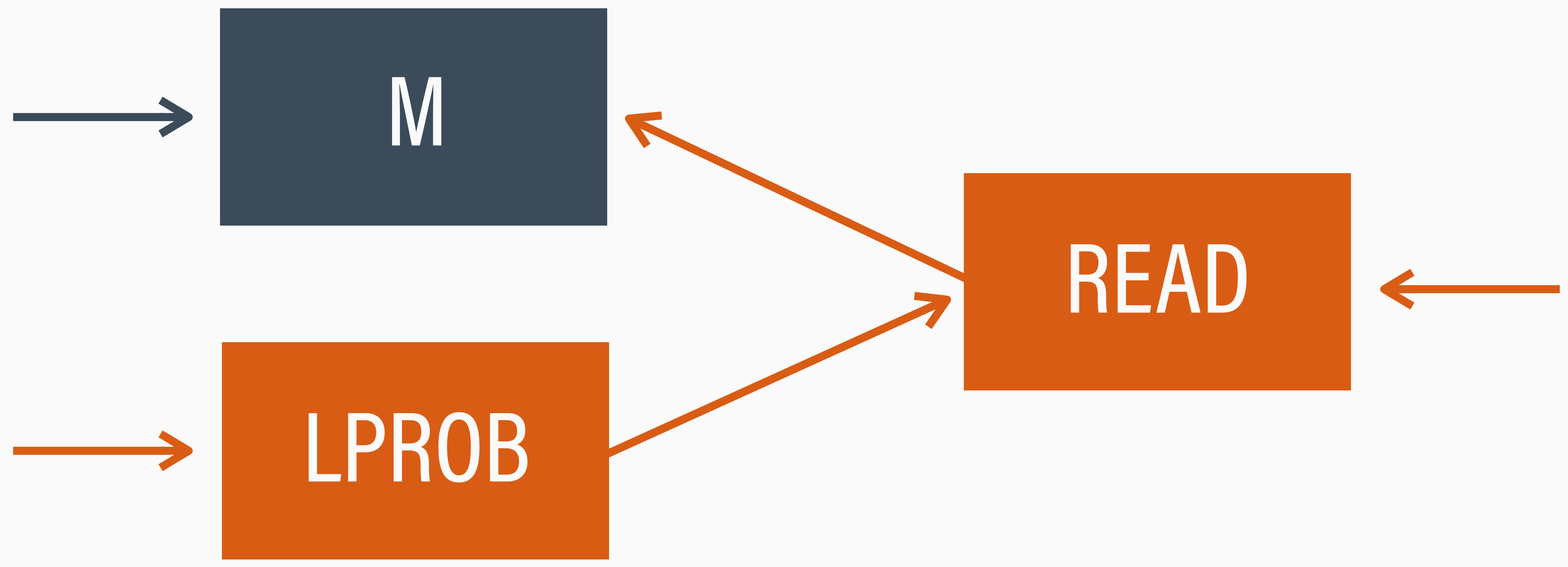
MNAR EXAMPLE

- Missingness is related to the observed learning problems measure and also related to the unseen reading scores
- Individuals with the low reading levels opt out because they feel discouraged or anxious about testing or because they were moved to specialized programs or alternative educational settings where standardized testing protocols differ

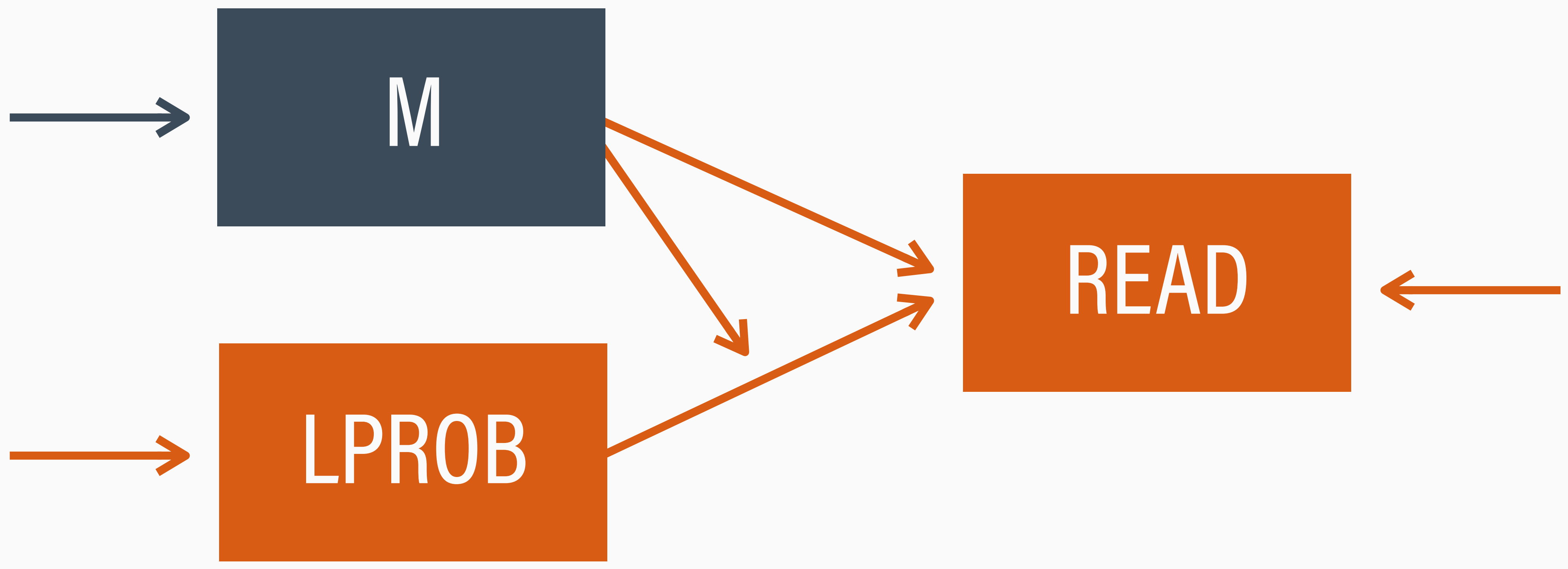
MNAR MODELING

- Missing not at random processes require an explicit model that incorporates the missing data indicator (M)

Selection Model



Pattern Mixture Model



TESTING THE CMAR ASSUMPTION

- The Big Three achieve unbiasedness if the process is conditionally MAR
- The CMAR assumption is untestable because it stipulates no relation between missingness and the unseen scores
- When in doubt, conduct sensitivity analyses that compare the estimates from CMAR and MNAR assumptions

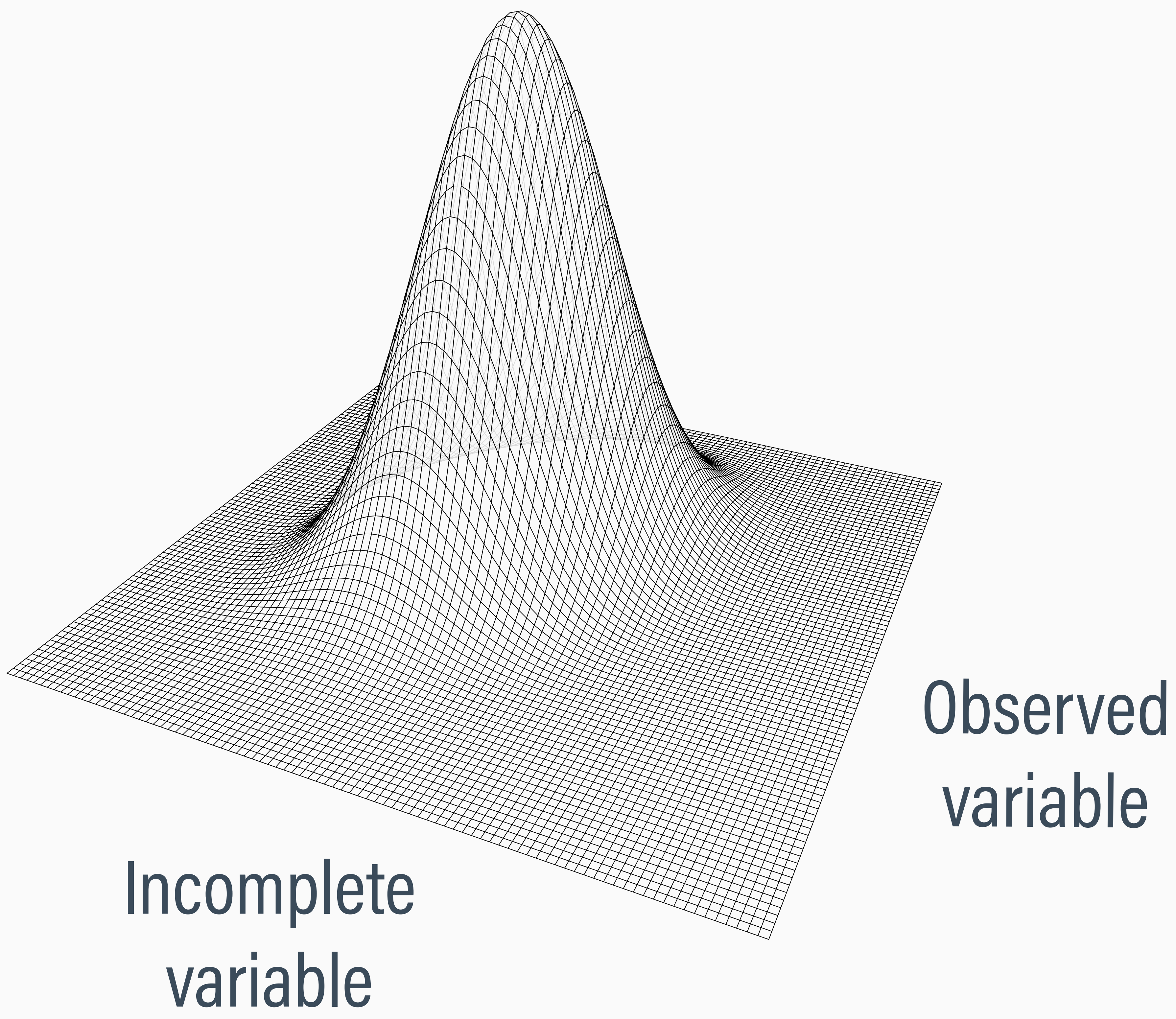
OUTLINE

MAXIMUM LIKELIHOOD

- ML identifies parameter estimates that minimize the distances between the model's predicted values and the observed data
- Each observation's contribution to estimation is restricted to the subset of parameters for which there is data
- Estimation uses incomplete data, no imputation performed

IMPLICIT IMPUTATION

- Each participant contributes their observed data
- Data are not filled in, but the multivariate normal distribution acts like an imputation machine
- The location of the observed data implies the probable position of the unseen data, and estimates are adjusted accordingly

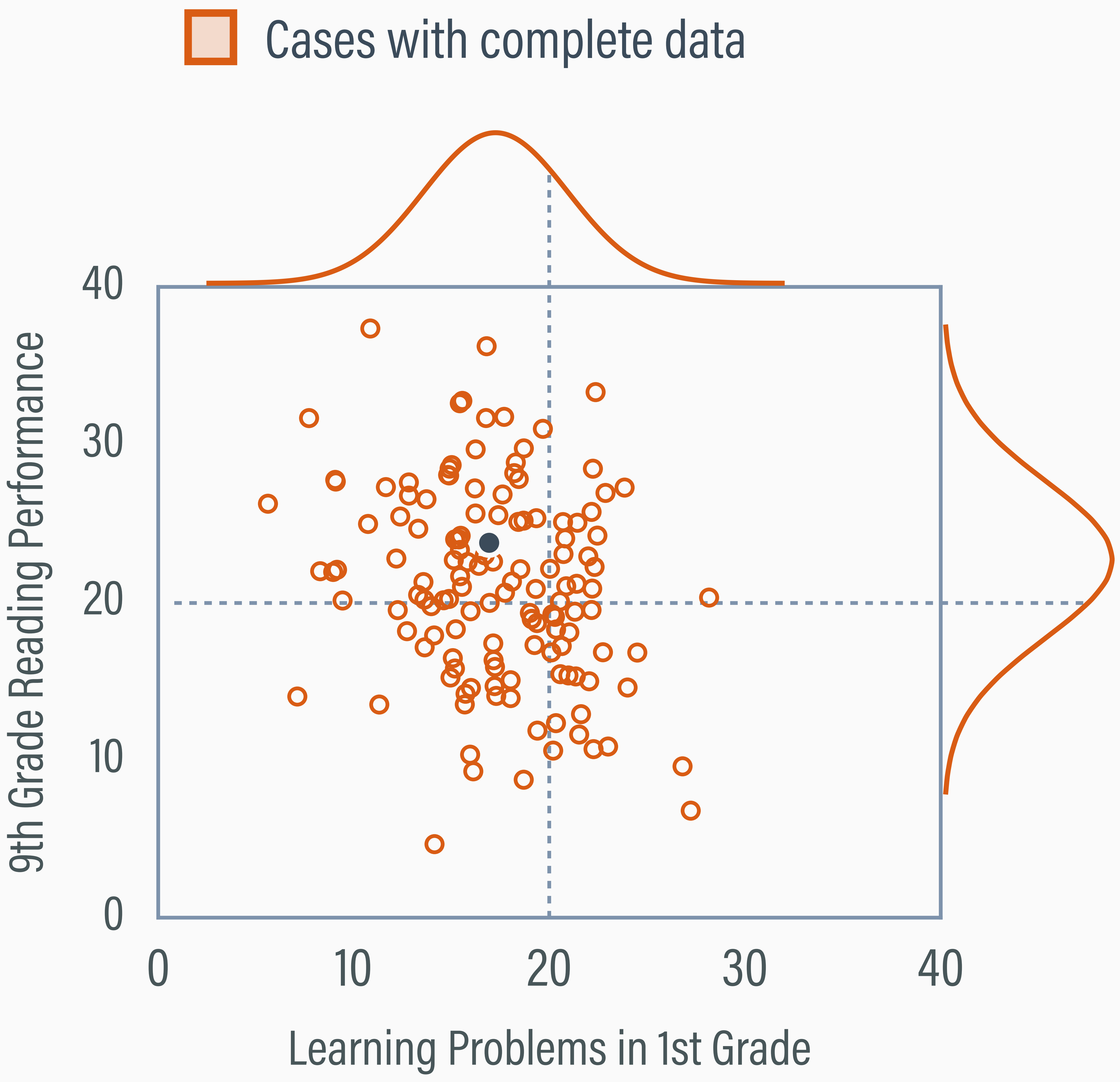


READING ILLUSTRATION

- Students with high learning problems ratings are more likely to have missing reading scores (conditionally MAR)
- The true means are both 20

DELETING INCOMPLETE DATA

- Deleting cases with missing reading scores gives a non-representative sample
- Scores are systematically missing from one side of each distribution
- The reading mean is too high and the learning problems mean is too low

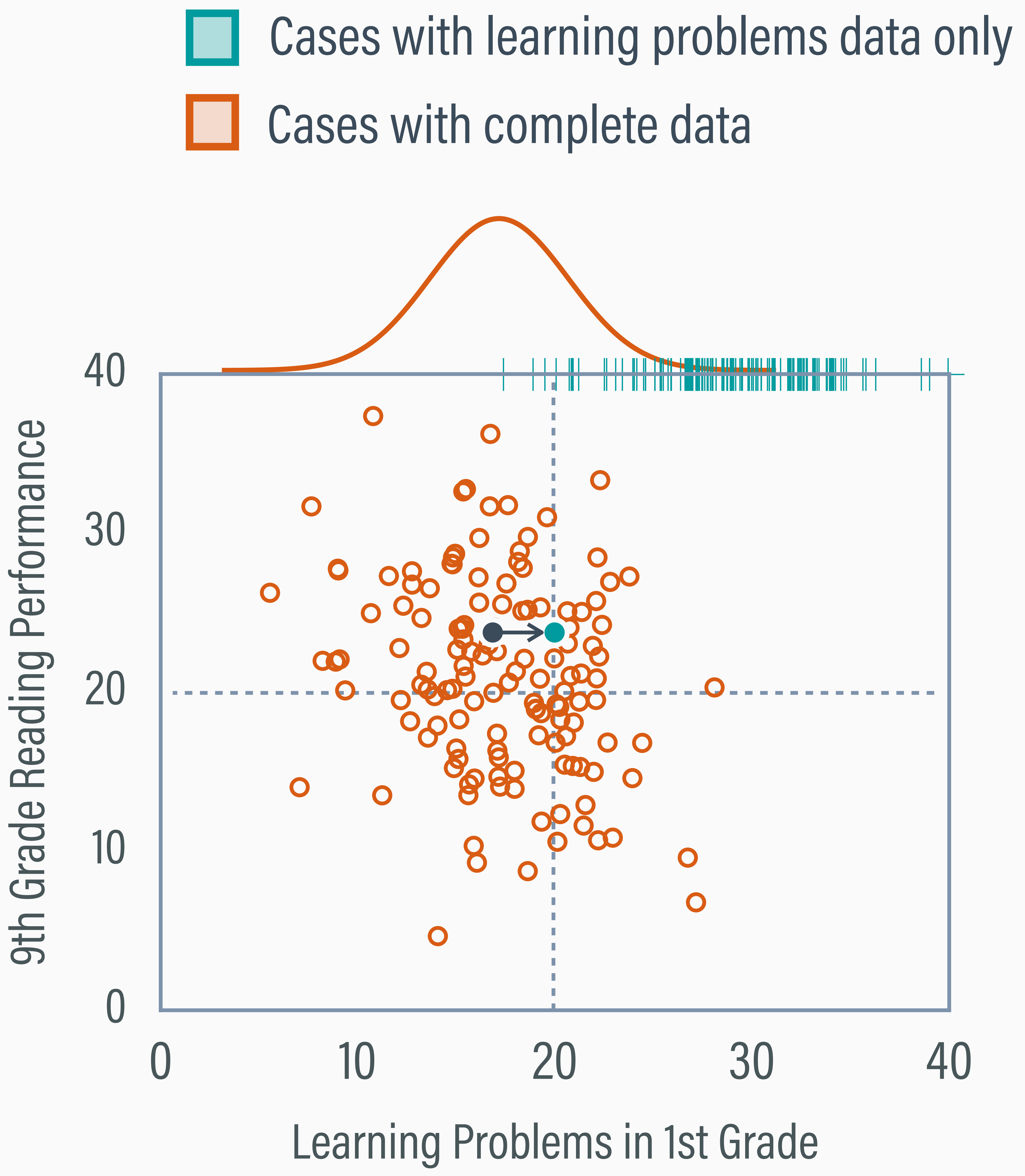


PARTIAL DATA RECORDS

- ML uses the partial data for students with learning problems data and missing reading scores
- The partial data records primarily have elevated learning problems ratings

ADJUSTING LEARNING PROBLEMS MEAN

- Adding higher learning problems scores increases the variable's variability
- The variable's mean receives a upward adjustment to accommodate the influx of high scores

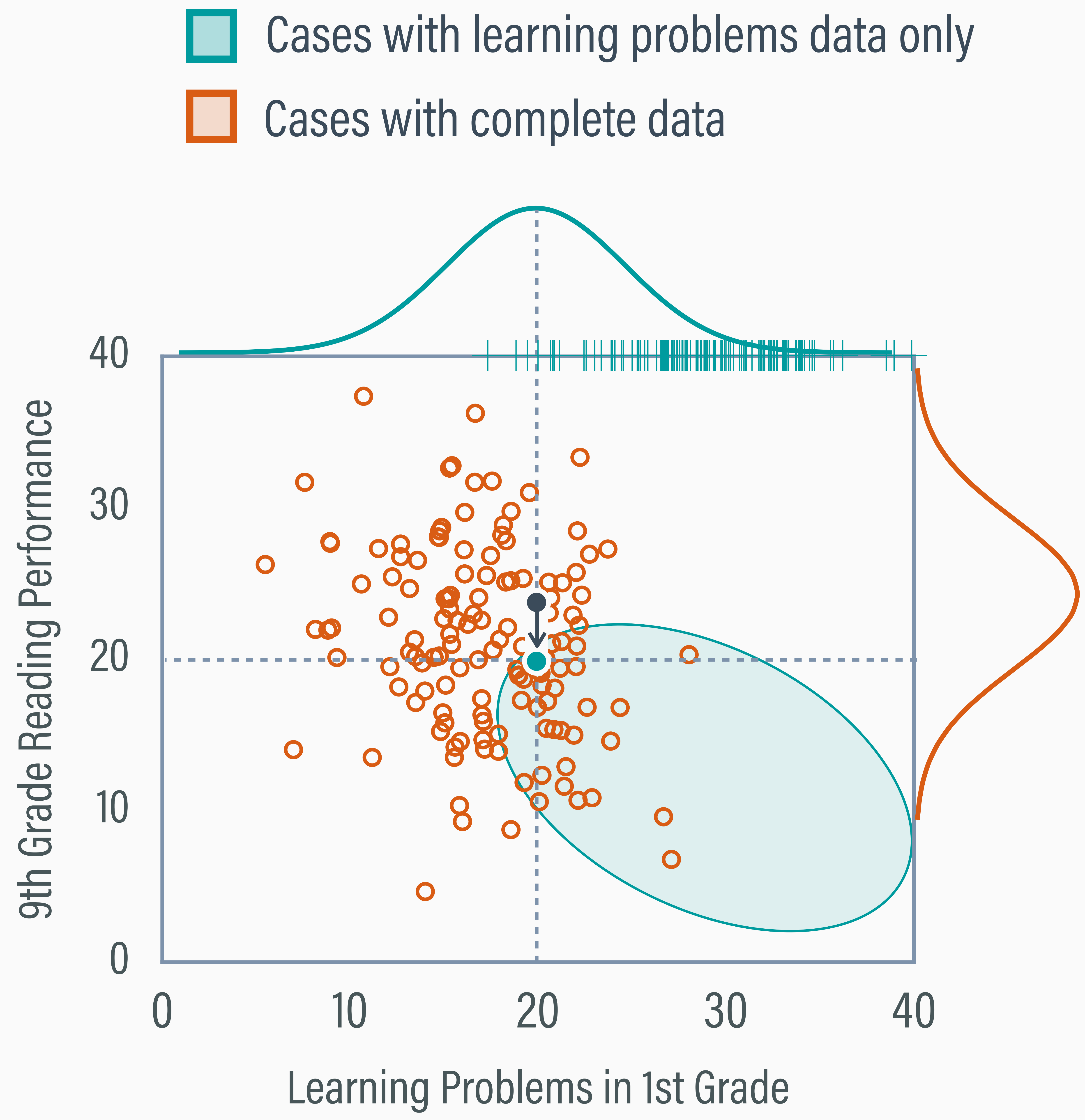


IMPLICIT IMPUTATION

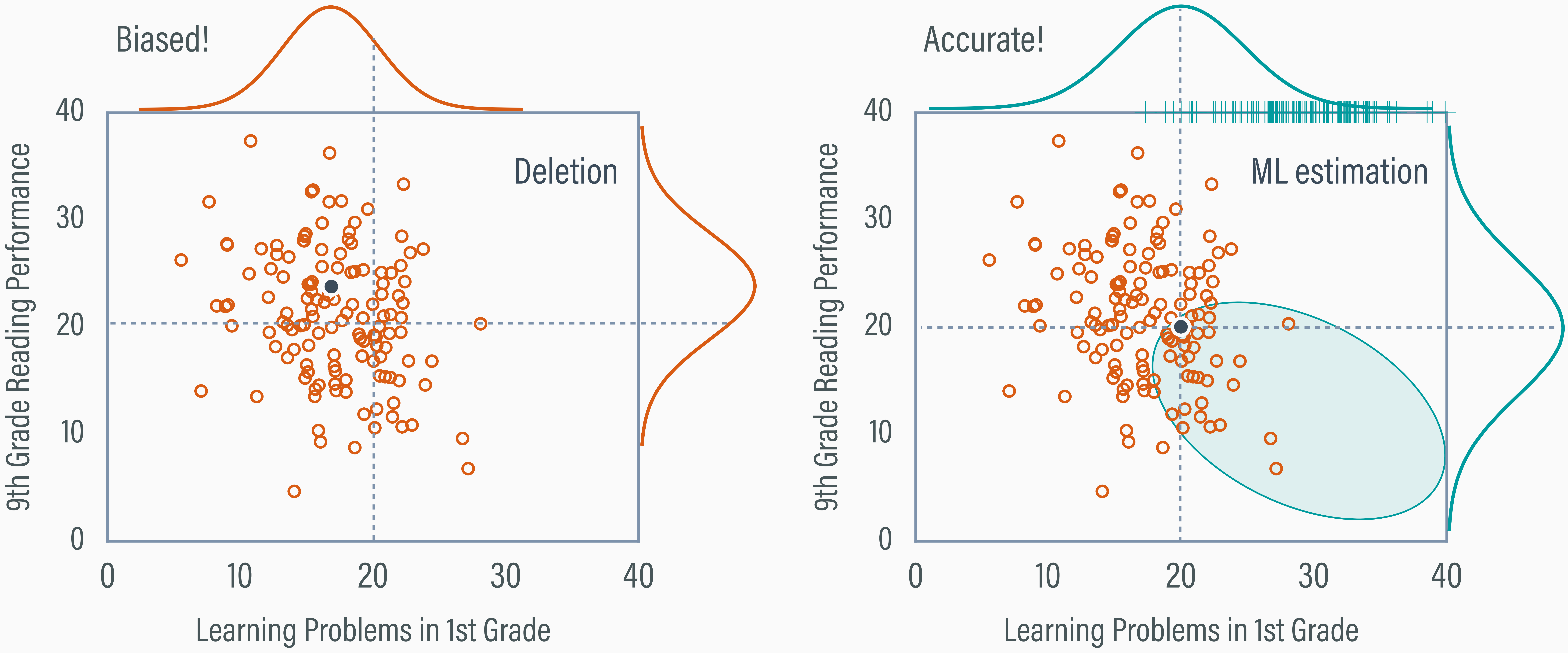
- Maximum likelihood assumes multivariate normality
- In a normal distribution with a negative correlation, higher learning problems scores should pair with lower reading

ADJUSTING THE READING DISTRIBUTION

- From the negative correlation, maximum likelihood intuits the presence of the lower but unseen reading scores
- The variance of the reading distribution increases, and the unseen reading scores in the lower tail adjust the mean down



ESTIMATION SUMMARY



MAXIMUM LIKELIHOOD PROS AND CONS

Pros	Cons
Direct estimation for a wide range of analysis models	Generally limited to normal data, options for mixed metrics are less common
Widely available in software packages (any SEM program)	Normal-theory methods are biased with interactions and non-linear terms
Easy to use, missing data handling occurs behind the scenes	MLM software usually discards observations with missing predictors

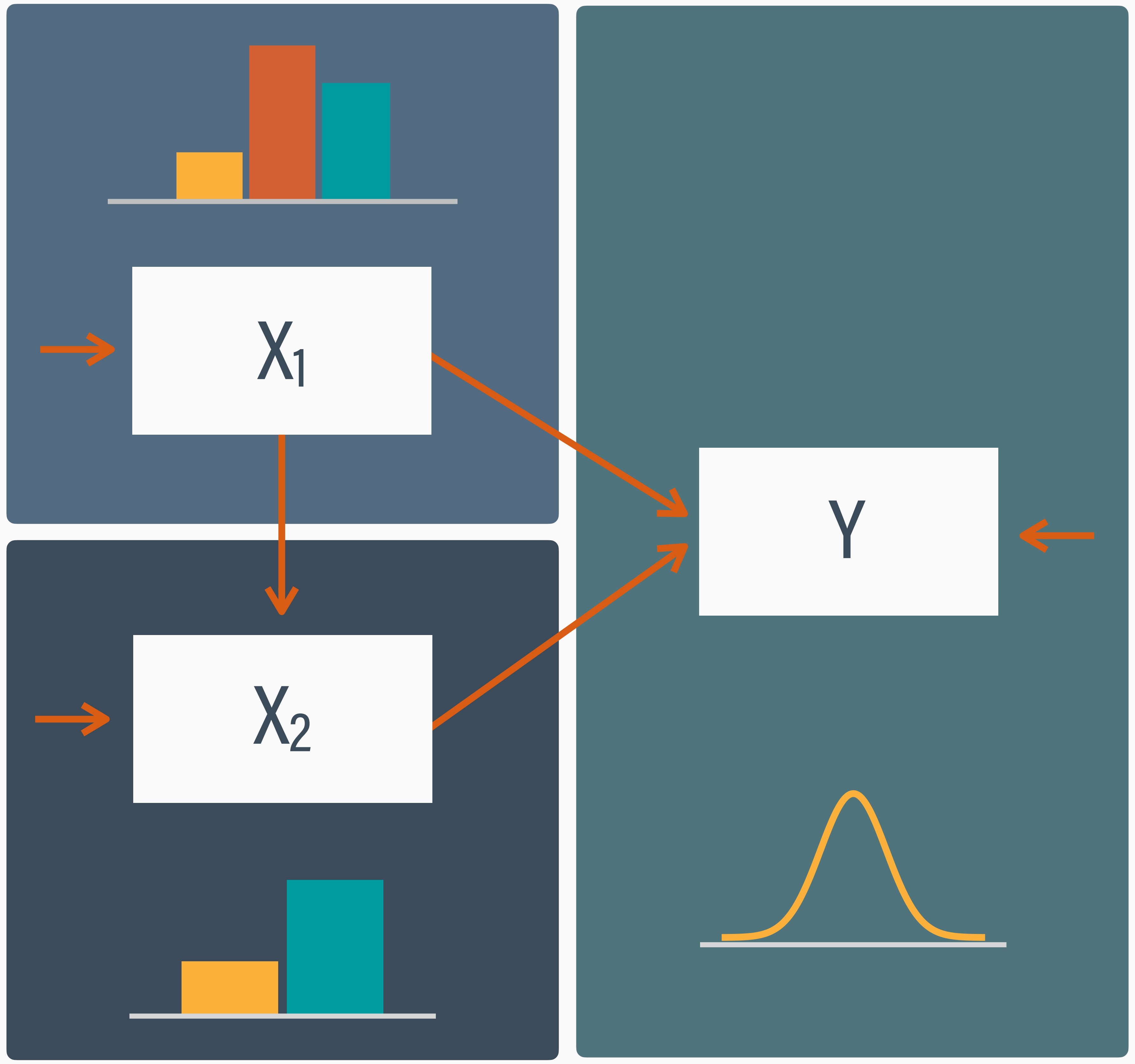
OUTLINE

THINGS MCMC ESTIMATION IS GOOD AT

- Direct estimation for complex models with missing data
- Mixed metrics (normal, ordinal, nominal, skewed, count, latent)
- Nonlinear effects (interactions, curvilinear effects)
- Multilevel data (random coefficients, interactions)
- Latent variable modeling (interactions)

FACTORED REGRESSION SPECIFICATIONS

- Factored regression specifications invoke a unique distribution for each variable
- The analysis consists of a collection of univariate regression models
- Each model can include terms that are at odds with multivariate normality



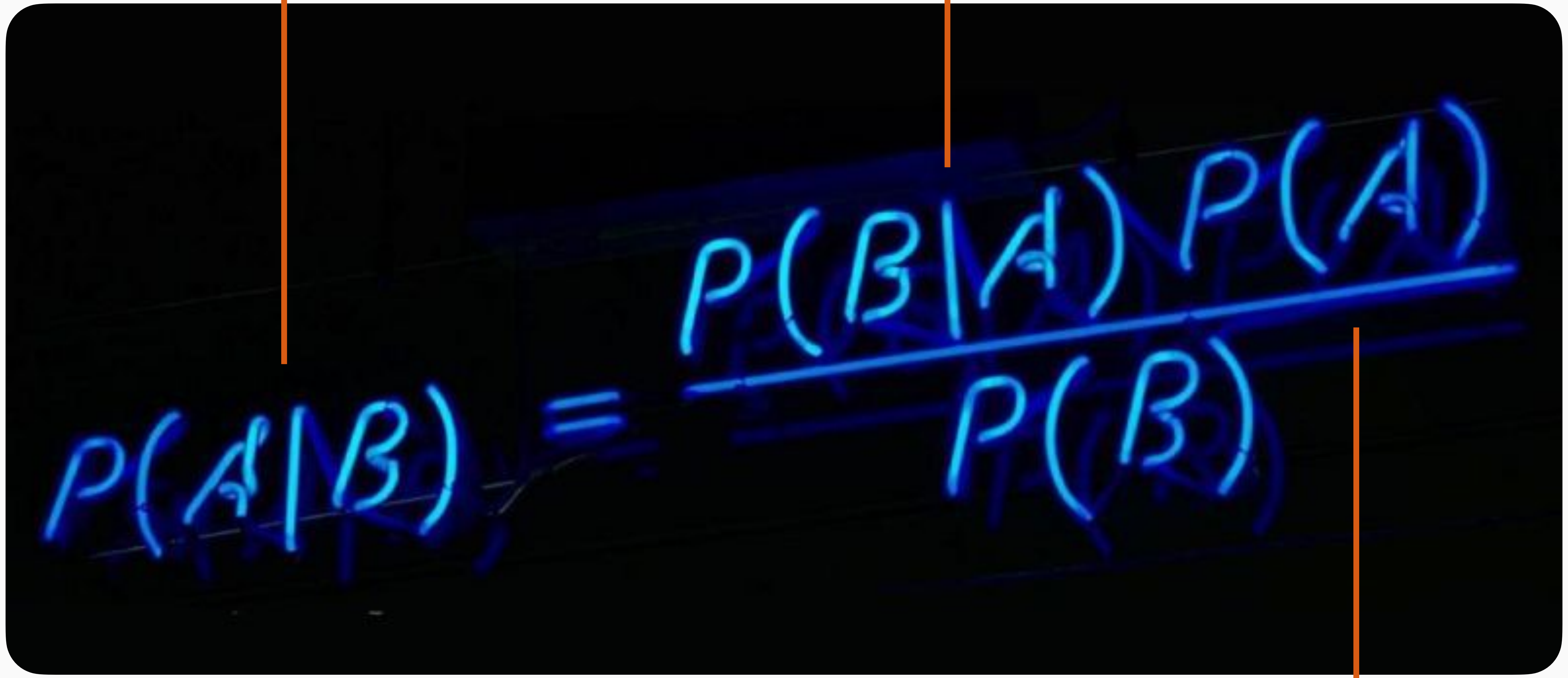
FREQUENTIST VS. BAYESIAN PARADIGMS

Frequentist	Bayesian
The parameter is a fixed quantity, estimates vary across different samples	Parameters are random variables with a distribution of plausible realizations
Statements about probability, precision, and confidence refer to estimates	Statements about probability, precision, and intervals refer to the parameter
Probability = long run frequency of outcomes across many samples	Probability = our degree of certainty about a parameter after analyzing data

BAYES' THEOREM

Posterior = parameters (A) given the data (B)

Likelihood = data (B) given the parameters (A)



Prior = a priori belief about parameters (A)

MCMC ESTIMATION

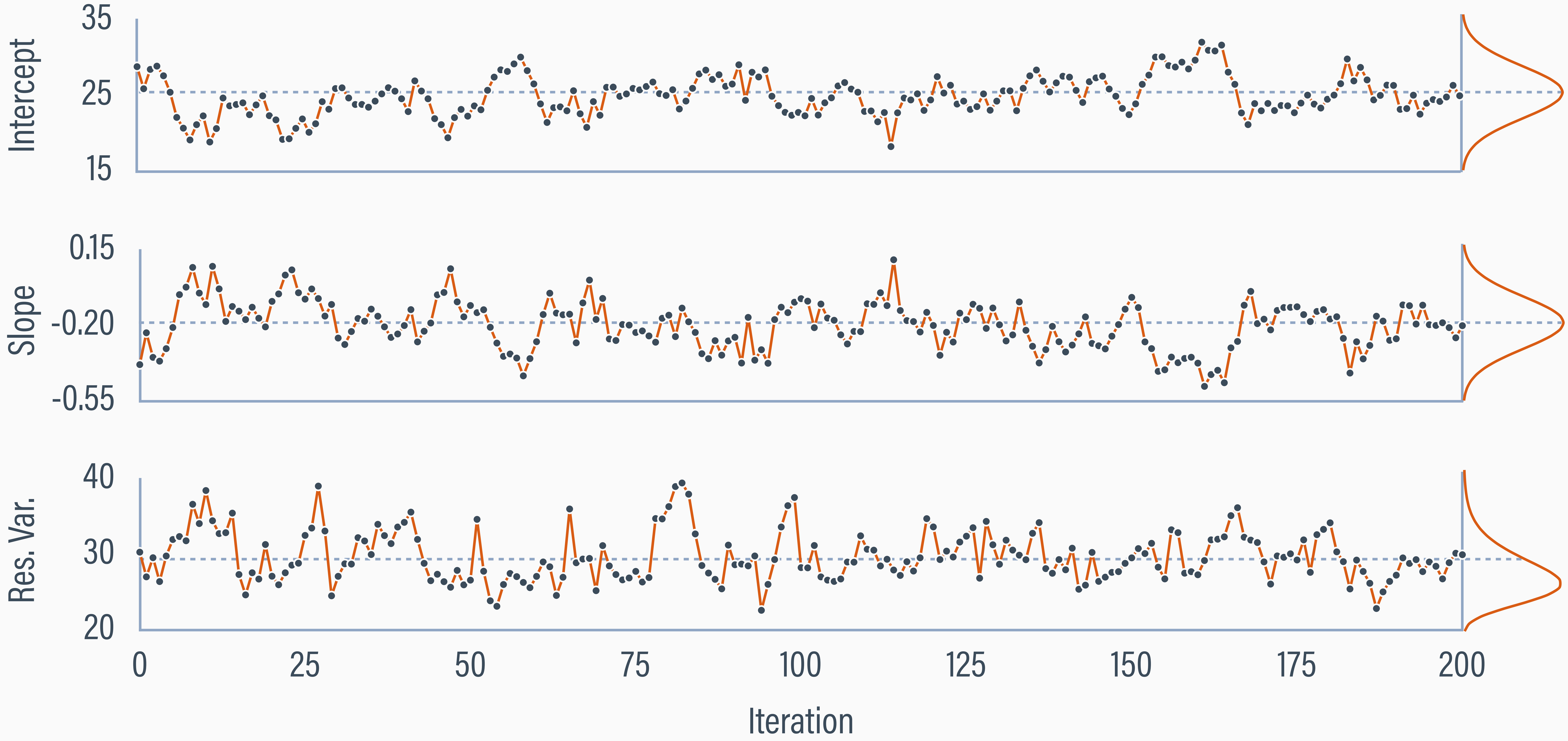
Do for t = 1 to 10,000 iterations

- Estimate model parameters, conditional on the filled-in data
- Impute missing values, conditional on the model parameters

Repeat

Summarize model parameters

PARAMETERS FROM 200 MCMC CYCLES

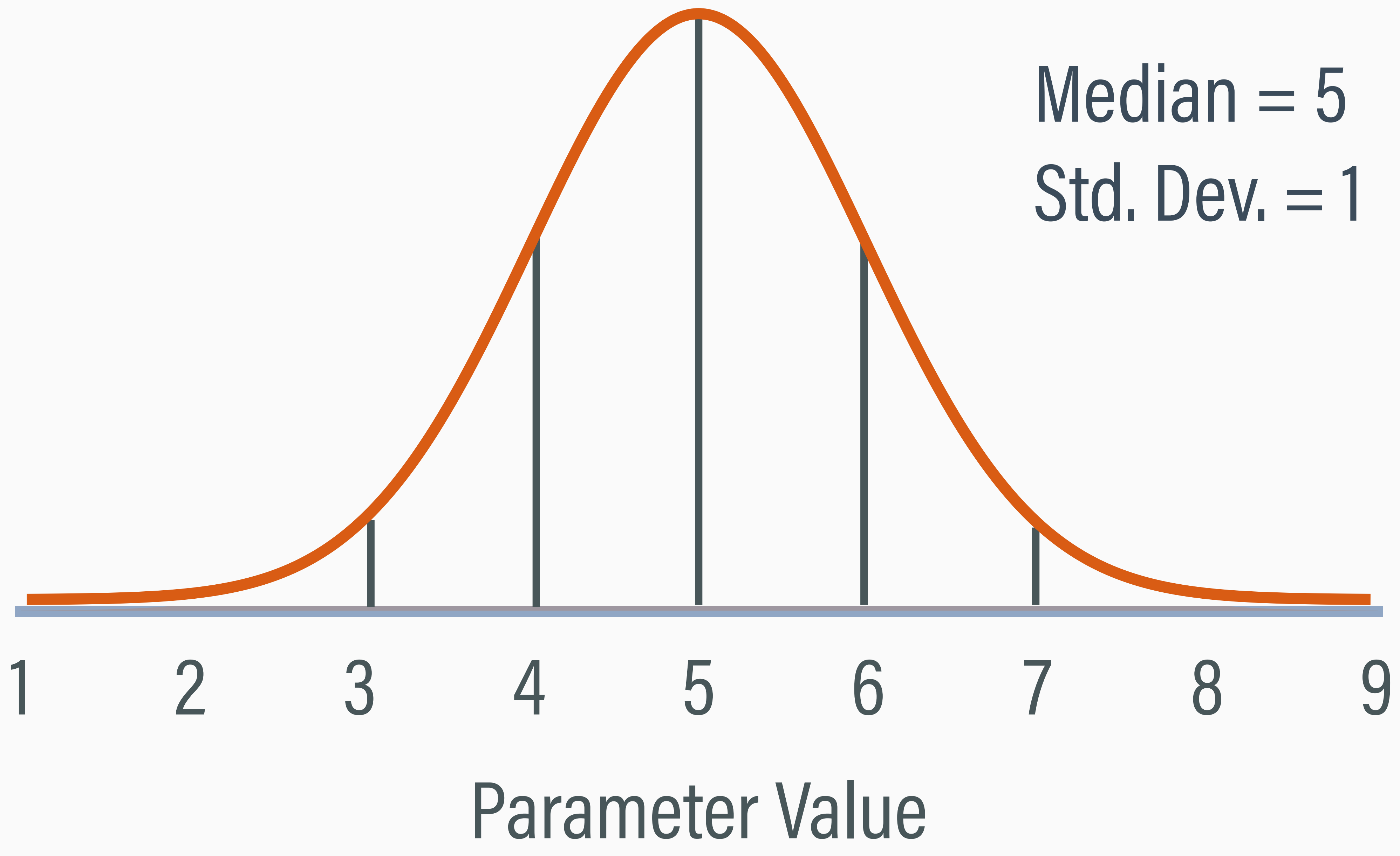


SUMMARIZING MCMC ESTIMATES

- Bayesian estimation yields a distribution of parameters—a posterior—that averages over thousands of filled-in data sets
- The posterior is a distribution of plausible parameter values that could have produced our particular data
- Some parameter values are more likely to have produced the observed data than others

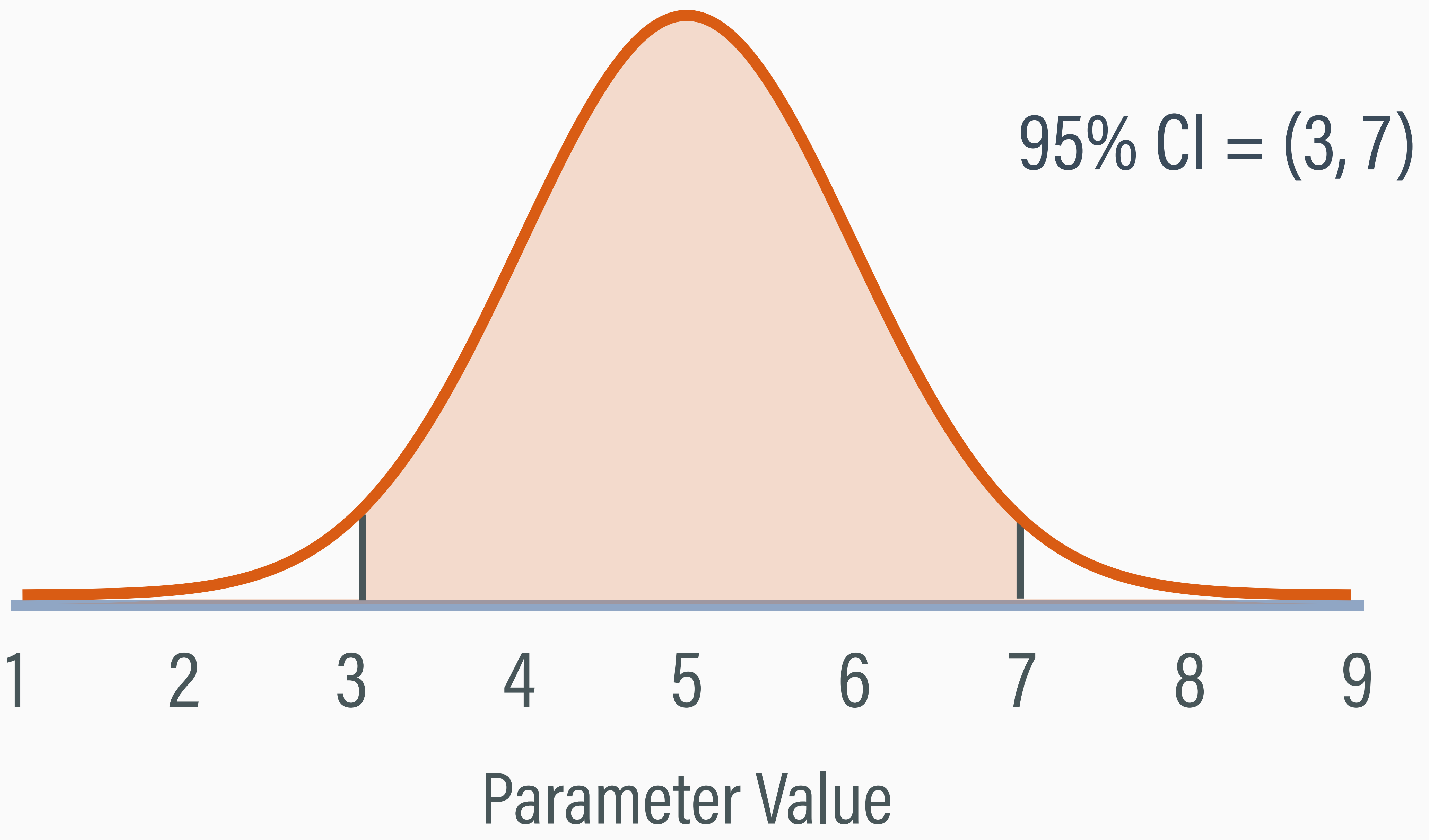
POSTERIOR MEDIAN AND STD. DEV.

- The posterior median and standard deviation quantify the most likely parameter value and uncertainty
- Analogous to a point estimate and standard error but no repeated sampling



95% CREDIBLE INTERVALS

- The 95% credible interval gives limits spanning 95% of the parameter’s range
- Akin to a confidence interval, but references a range of highly plausible parameter values



PHILOSOPHICAL SPECTRUM

Bayesian Inference

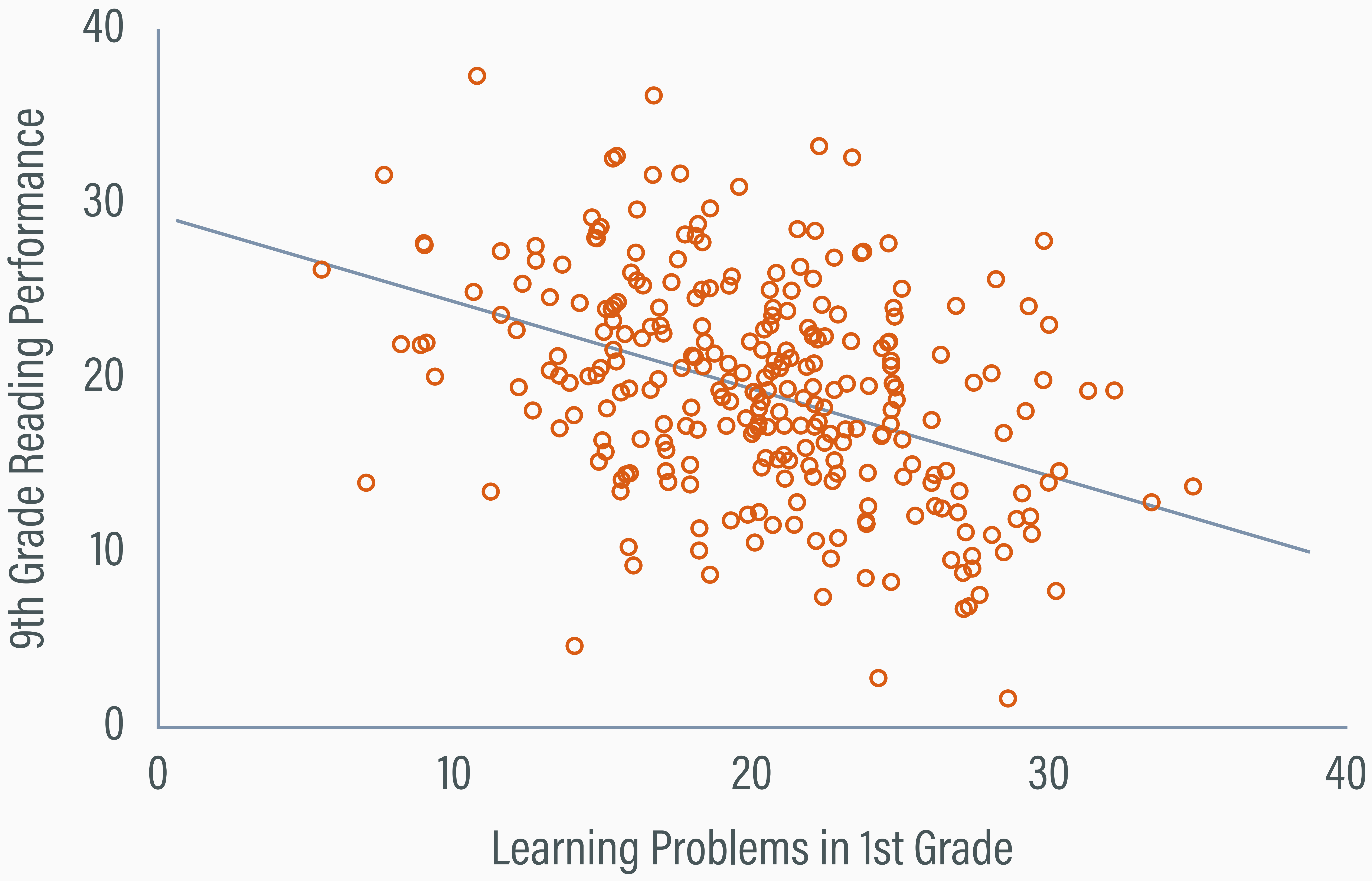
Computational frequentism

Computational frequentism: A researcher employs Bayesian MCMC estimation because their model is too complex for ML or OLS. The MCMC results are surrogates for unobtainable ML/OLS estimates (Levy & McNeish, 2021, Psychological Methods).

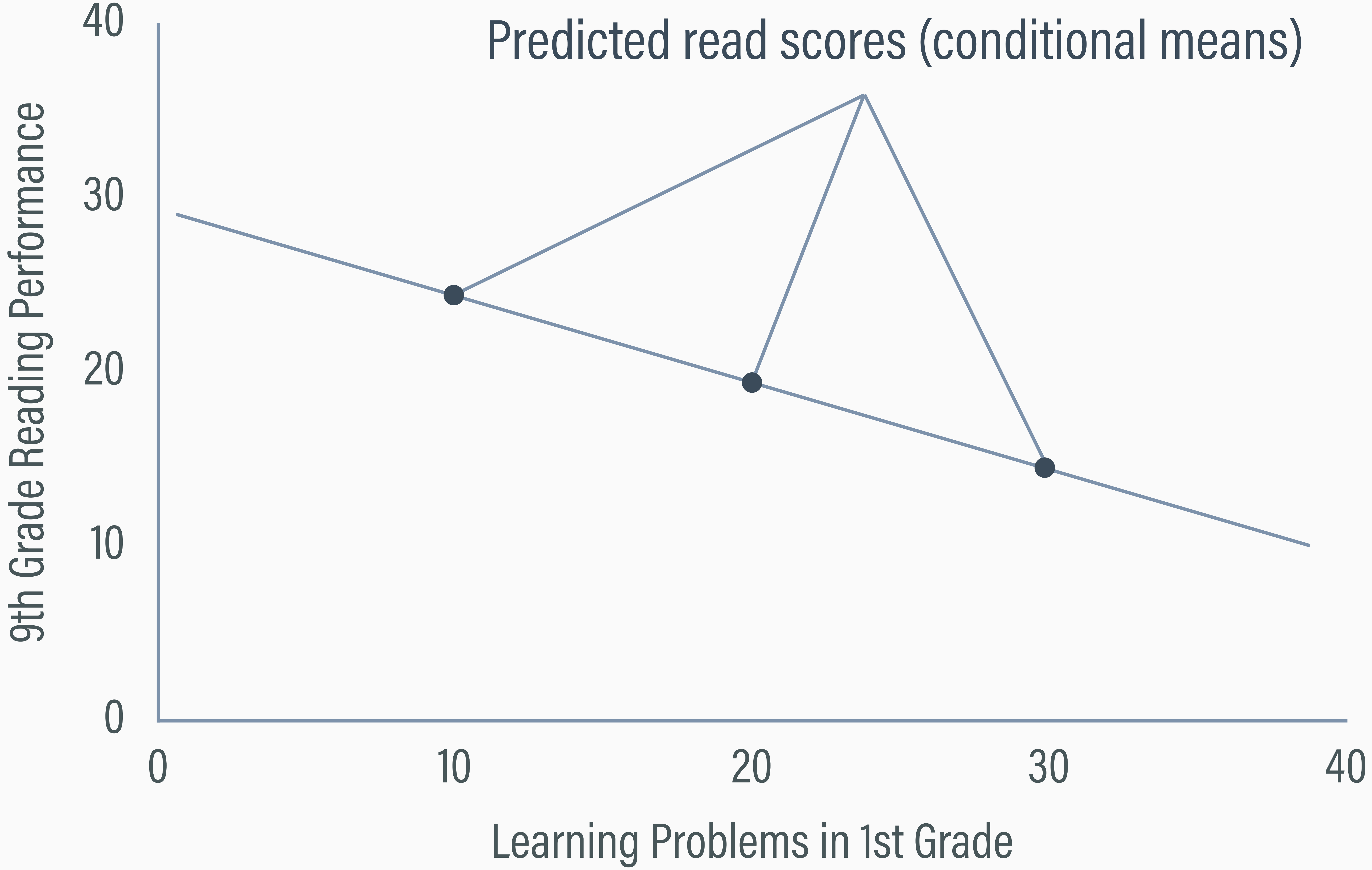
MISSING DATA IMPUTATION

- Missing scores are imputed by drawing replacement scores at random from a distribution of plausible values
- The model parameters combine to define the center and spread of the missing data imputations
- Each iteration yields unique model parameters and unique imputations based on those parameters

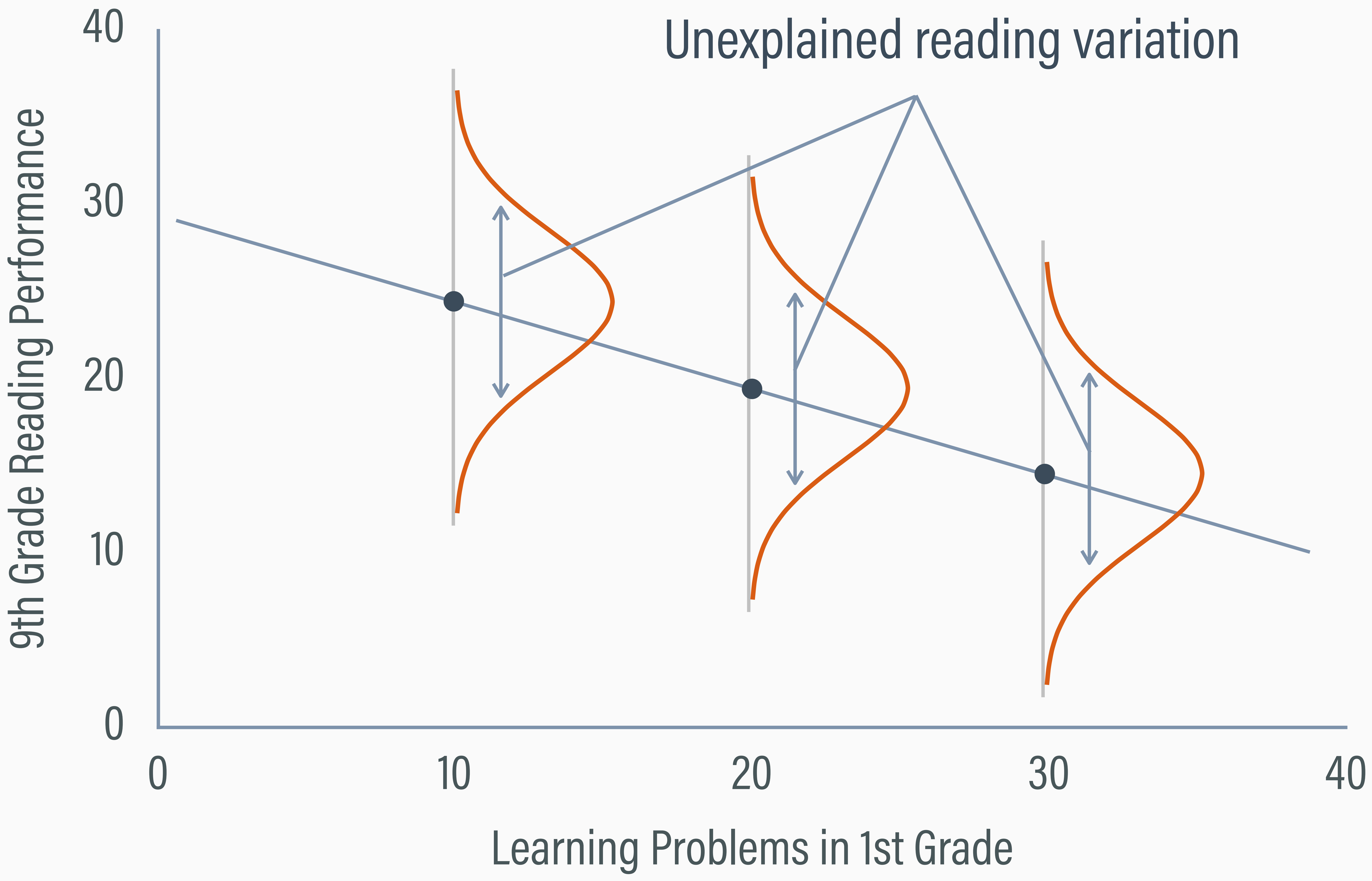
REGRESSION FROM FILLED-IN DATA



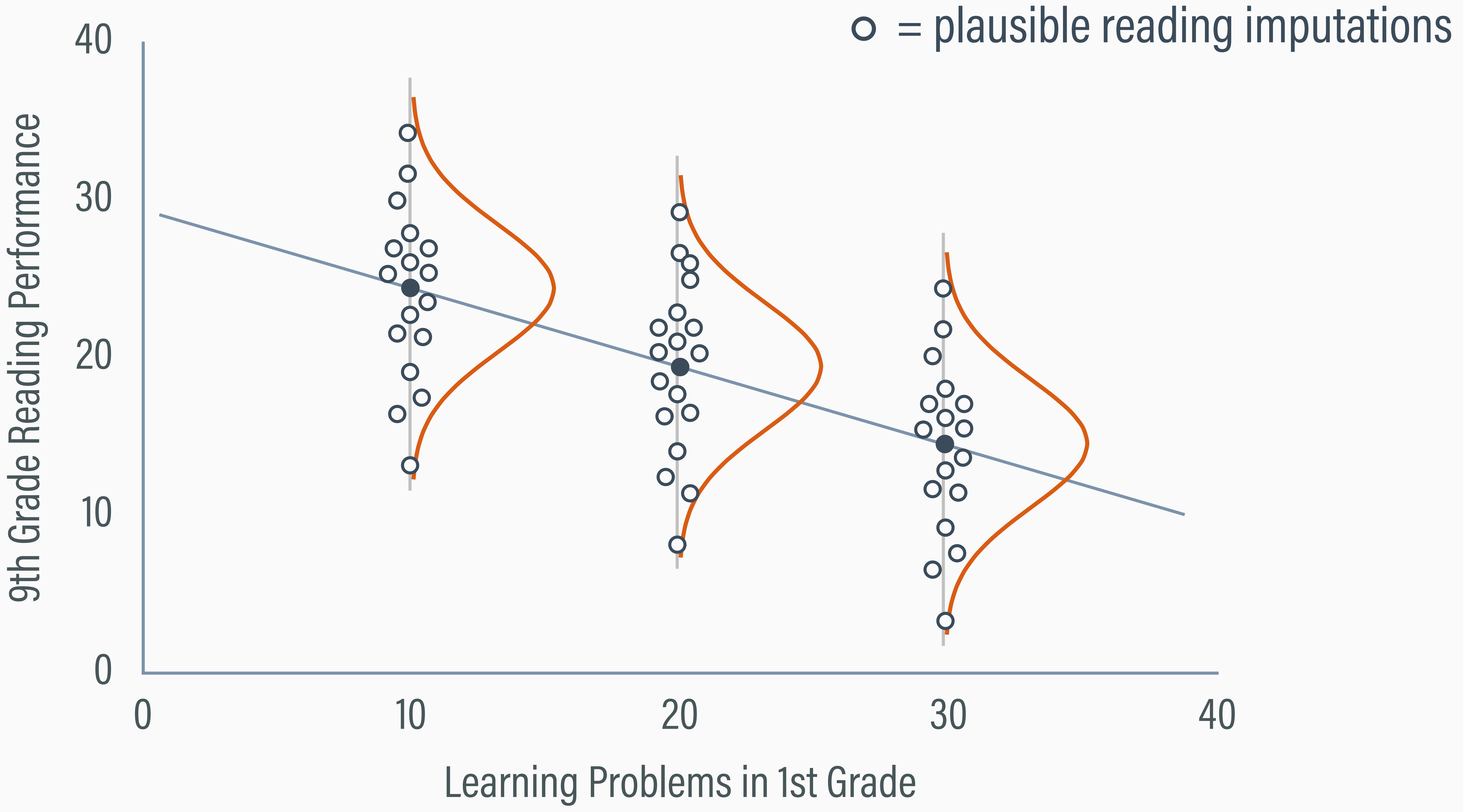
PREDICTED VALUES



RESIDUAL VARIATION

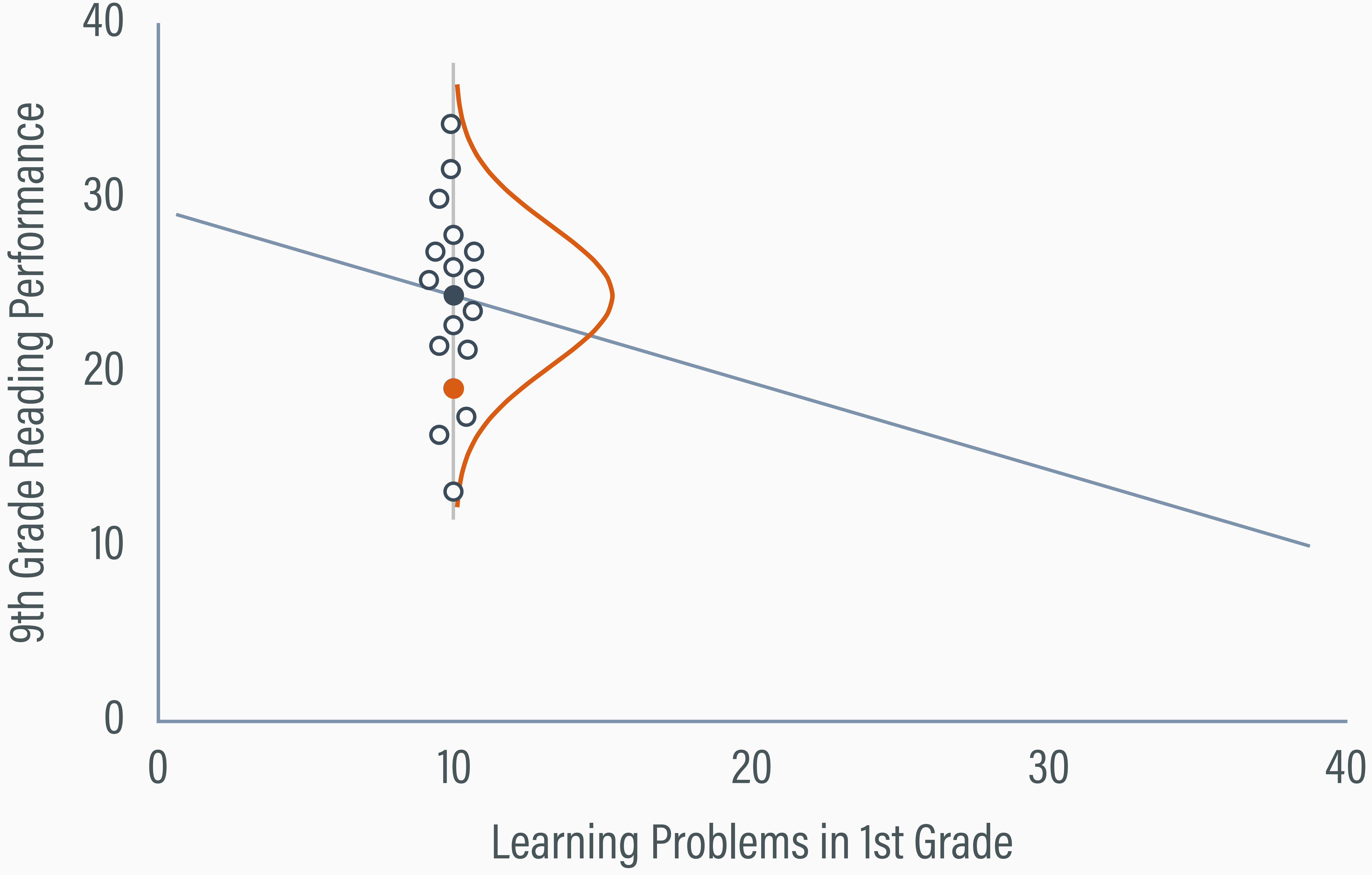


DISTRIBUTIONS OF IMPUTATIONS



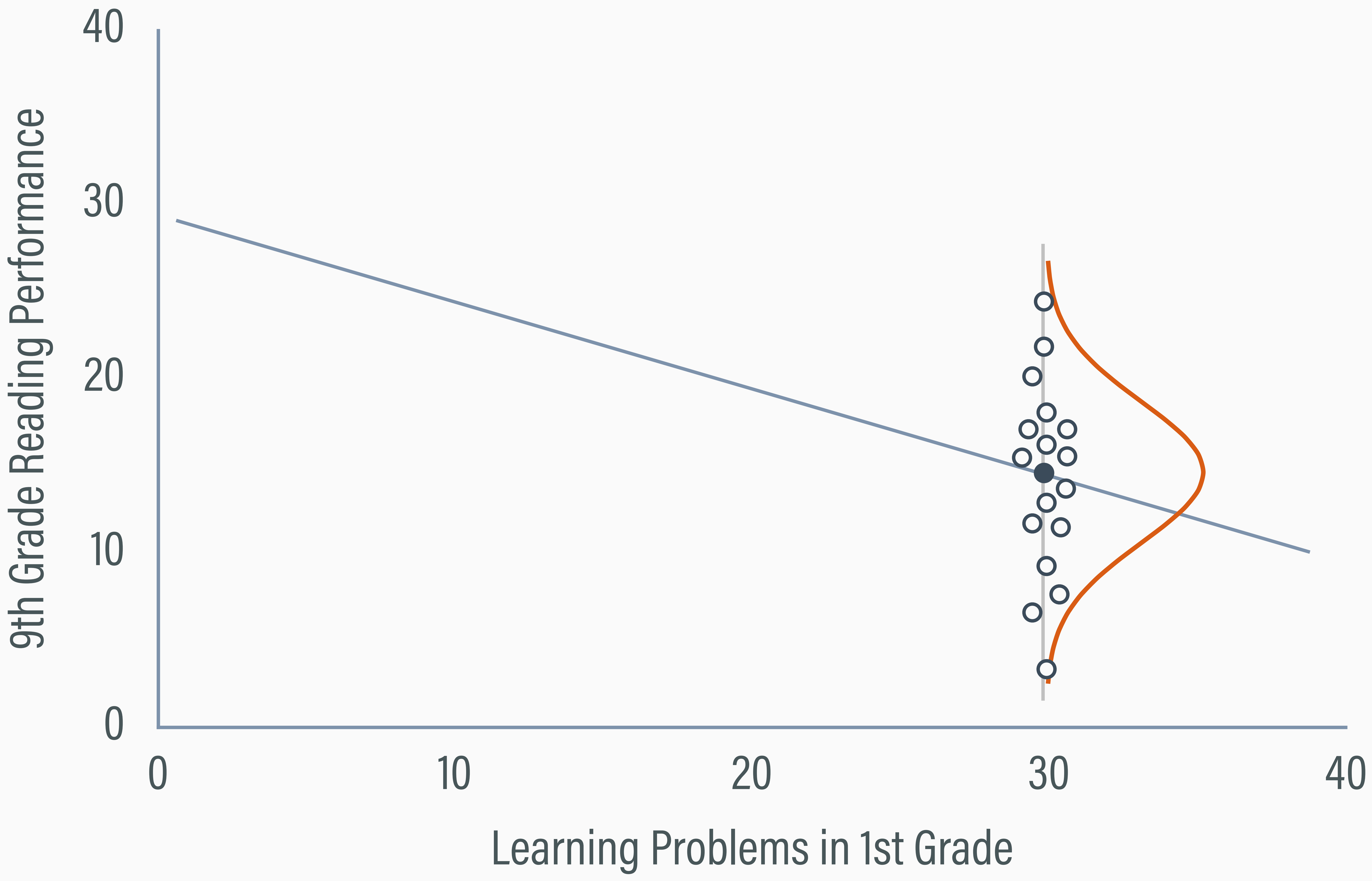
IMPUTATION FOR LOW LEARNING PROBLEMS

DRAW AN IMPUTATION AT RANDOM

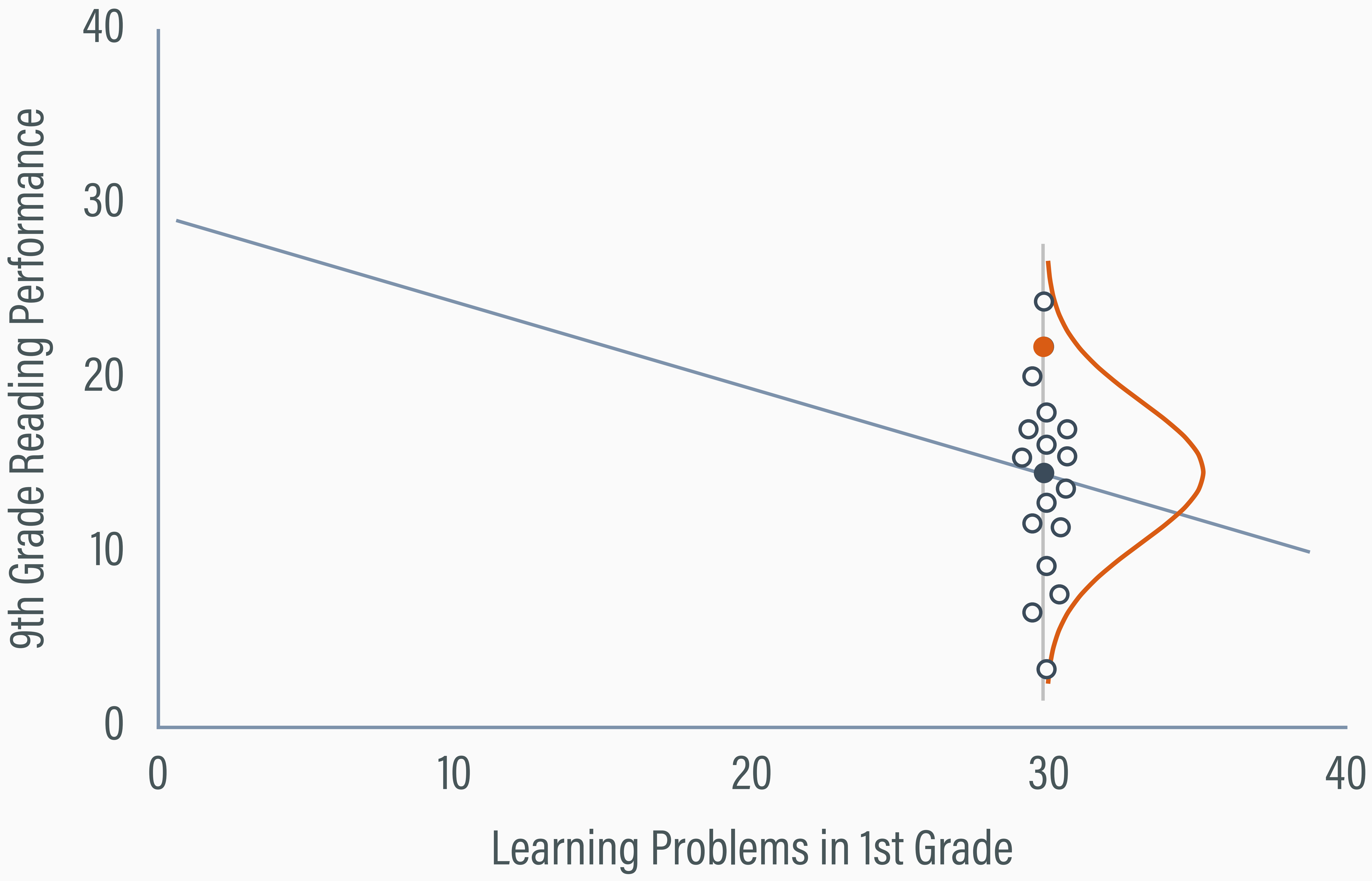


IMPUTATION = PREDICTION + NOISE

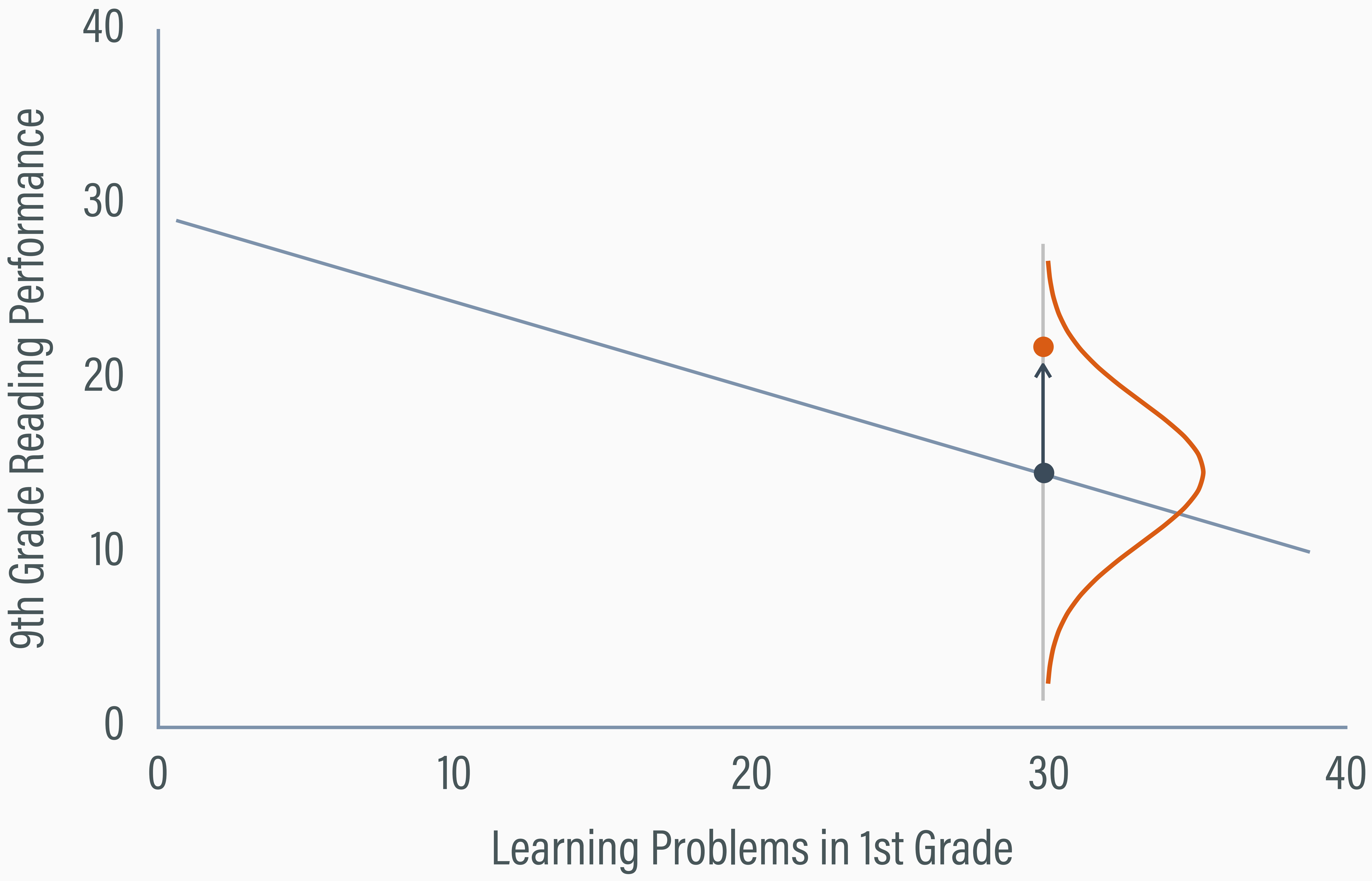
IMPUTATION FOR HIGH LEARNING PROBLEMS



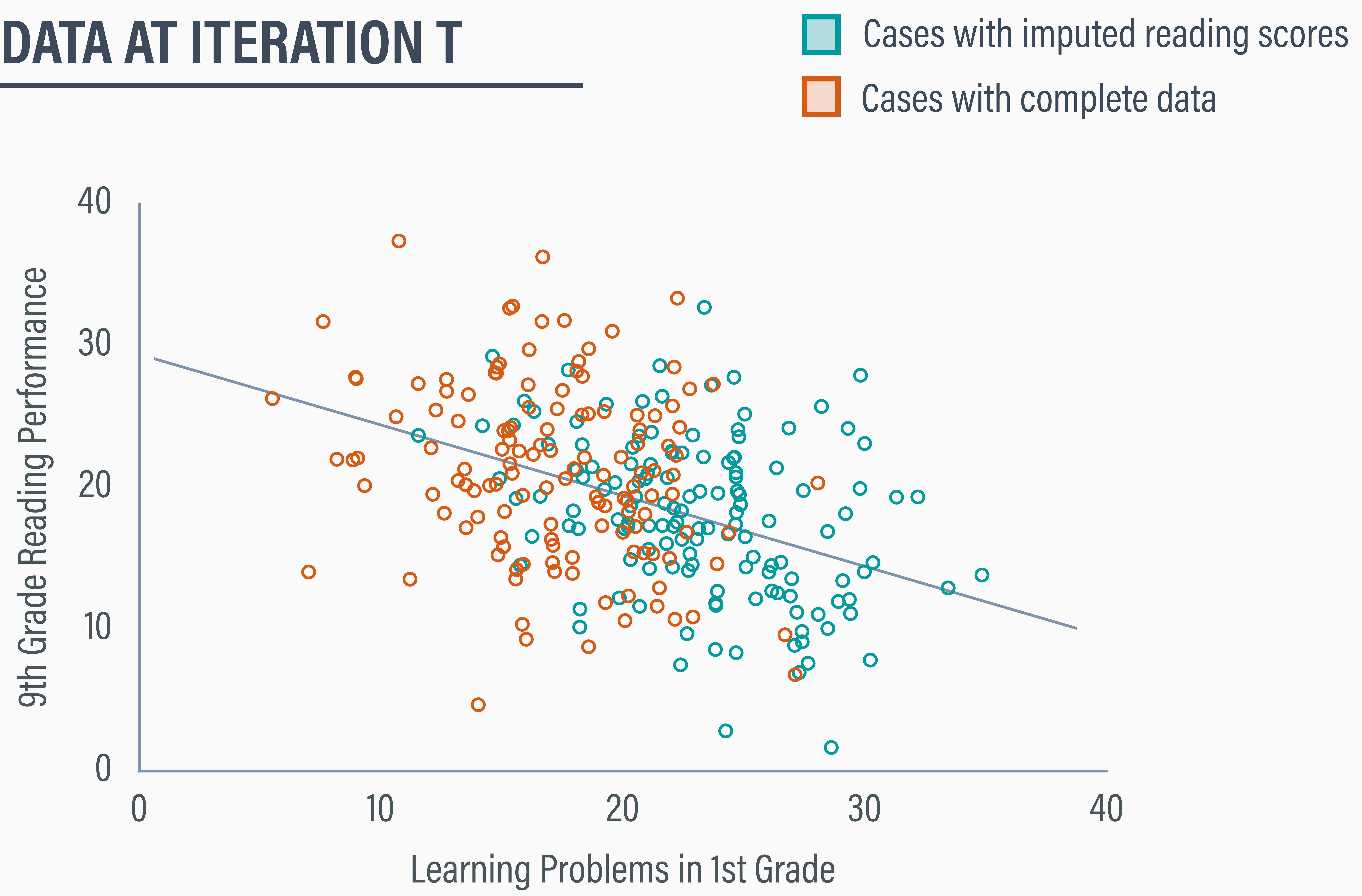
DRAW AN IMPUTATION AT RANDOM



DRAW AN IMPUTATION AT RANDOM



FILLED-IN DATA AT ITERATION T



BAYESIAN ESTIMATION PROS AND CONS

Pros

- Direct estimation competitor to maximum likelihood, but more flexible
- Suited for interactions, non-linear terms, and random coefficients (MLMs)
- Good for mixed metrics (normal, binary, ordinal, multicategorical, count, skewed)

Cons

- Fewer simple software options (Blimp), most are difficult to use (JAGS)
- MCMC is not fully autonomous, requires some input and oversight
- Literature on factored regression specifications is less mature

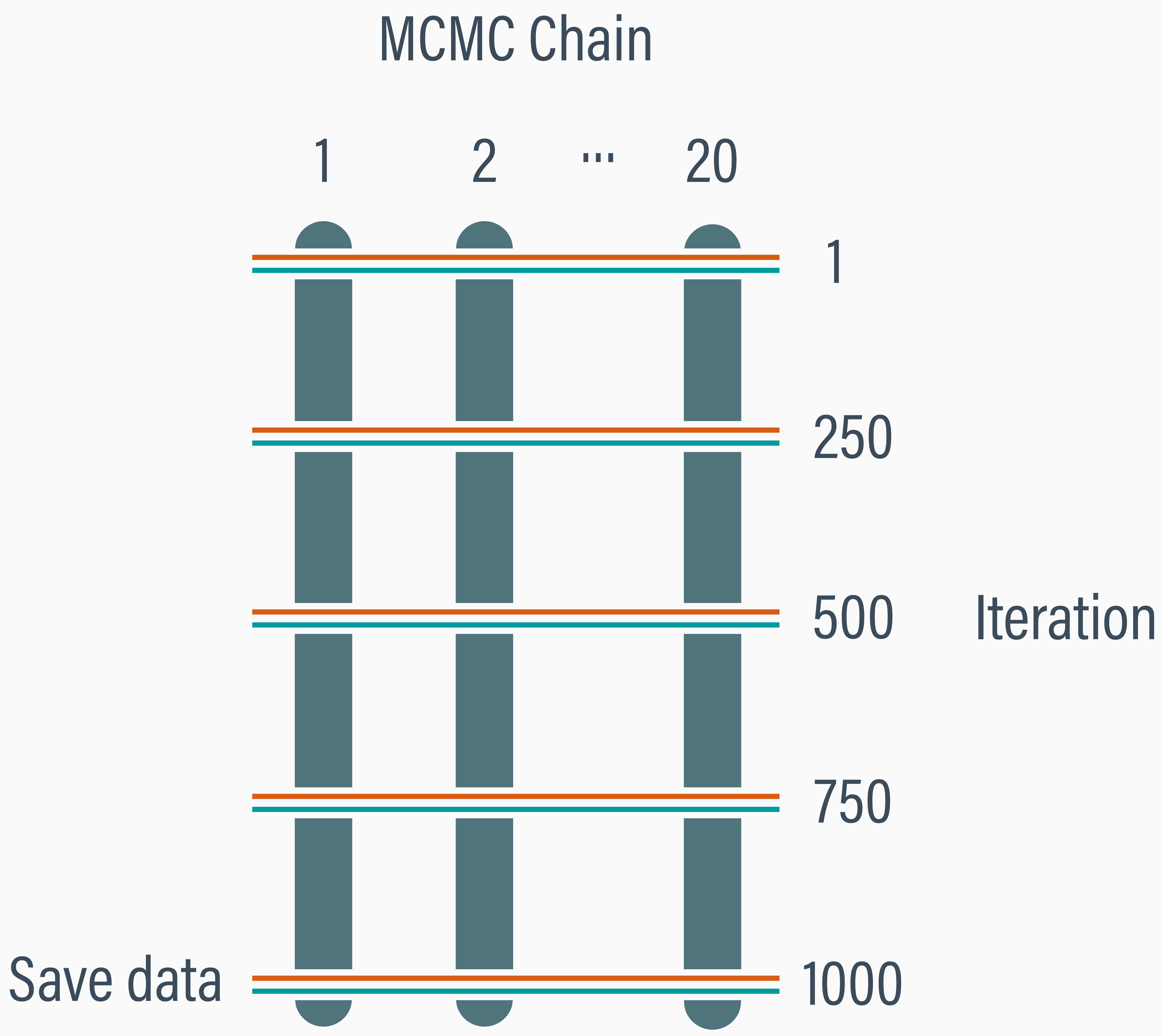
OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Bayesian MCMC Estimation
5. **Multiple Imputation**
6. Missing Data Software
7. Analysis Example

MULTIPLE IMPUTATION

- MCMC estimation creates a filled-in data set at every iteration, and estimates average over thousands of imputations
- Multiple imputation saves a small number of data sets (e.g., 20 is common) for reanalysis using frequentist methods
- MCMC is co-opted for the purpose of creating imputations, but the Bayesian parameter estimates are not of interest

MCMC ESTIMATION



STEP 1: SAVE IMPUTED DATA SETS

Original data

Y	X_1	X_2
4	4	3
3	NA	5
7	1	6
NA	1	6
5	9	3
3	NA	NA
1	6	7
9	4	9
2	NA	6

Imputed data set 1

Y	X_1	X_2
4	4	3
3	3.2	5
7	1	6
5.3	1	6
5	9	3
3	8.7	10.1
1	6	7
9	4	9
2	6.5	6

Imputed data set 2

Y	X_1	X_2
4	4	3
3	5.4	5
7	1	6
6.2	1	6
5	9	3
3	7.1	8.5
1	6	7
9	4	9
2	6.9	6

...

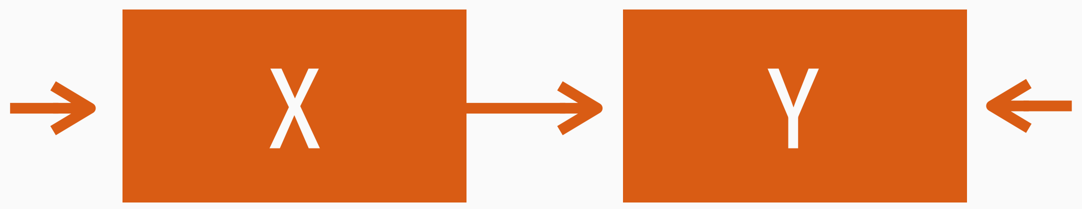
Imputed data set 20

Y	X_1	X_2
4	4	3
3	5.1	5
7	1	6
4.6	1	6
5	9	3
3	10.3	6.9
1	6	7
9	4	9
2	7.2	6

STEP 2: ANALYZE EACH DATA SET

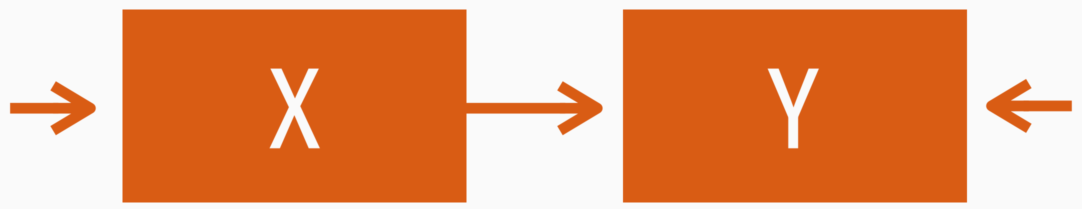
Analyze data set 1

Y	X_1	X_2
4	4	3
3	3.2	5
7	1	6
5.3	1	6
5	9	3
3	8.7	10.1
1	6	7
9	4	9
2	6.5	6



Analyze data set 2

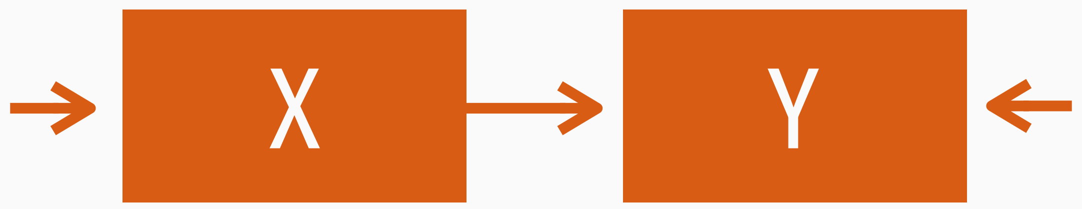
Y	X_1	X_2
4	4	3
3	5.4	5
7	1	6
6.2	1	6
5	9	3
3	7.1	8.5
1	6	7
9	4	9
2	6.9	6



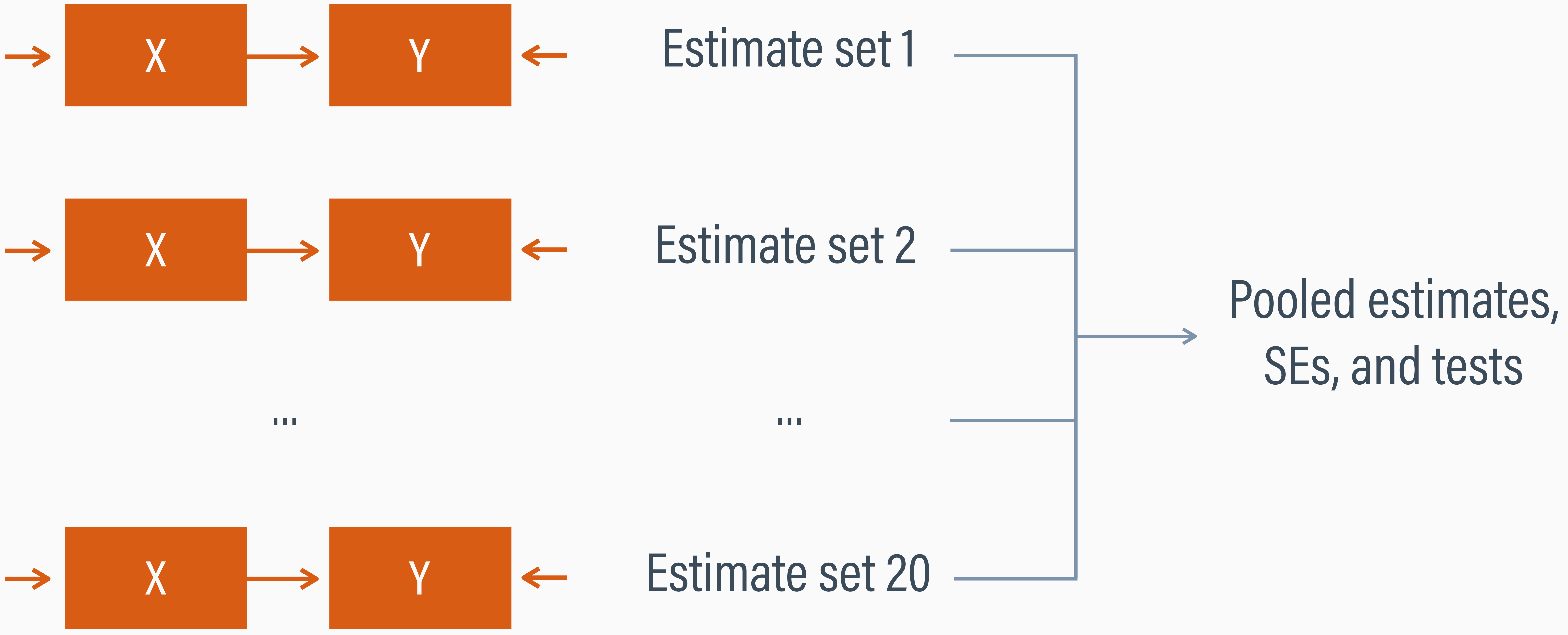
. . .

Analyze data set 20

Y	X_1	X_2
4	4	3
3	5.1	5
7	1	6
4.6	1	6
5	9	3
3	10.3	6.9
1	6	7
9	4	9
2	7.2	6



STEP 3: POOL RESULTS



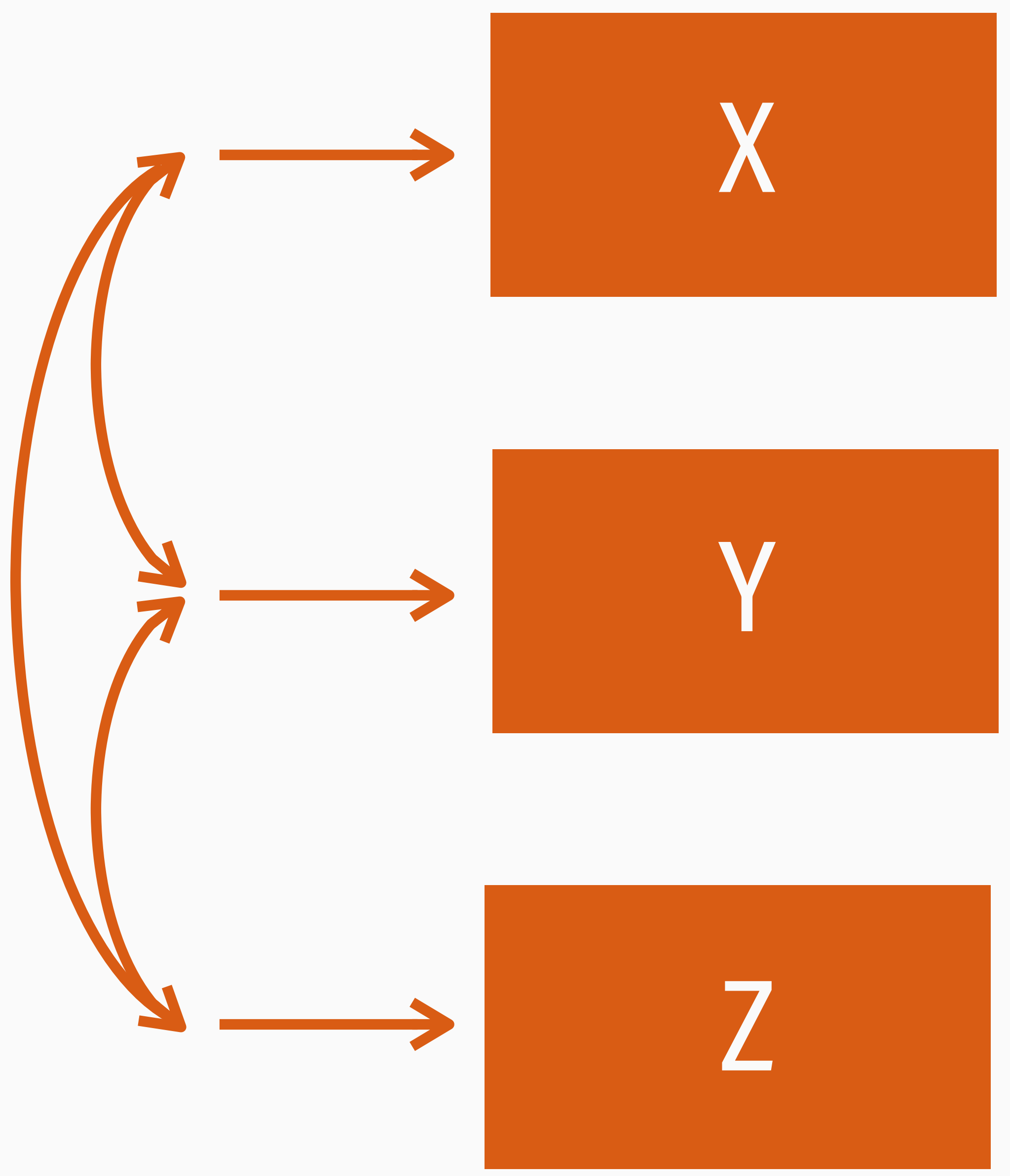
AGNOSTIC VS. MODEL-BASED IMPUTATION

- Step 1 (imputation) uses MCMC to fit a model, the parameters of which define distributions of imputations
- Step 2 (analysis) fits the focal models to the filled-in data
- Agnostic imputation deploys an imputation model that differs from the analysis model, whereas model-based imputation deploys the same model in both steps

JOINT MODEL IMPUTATION

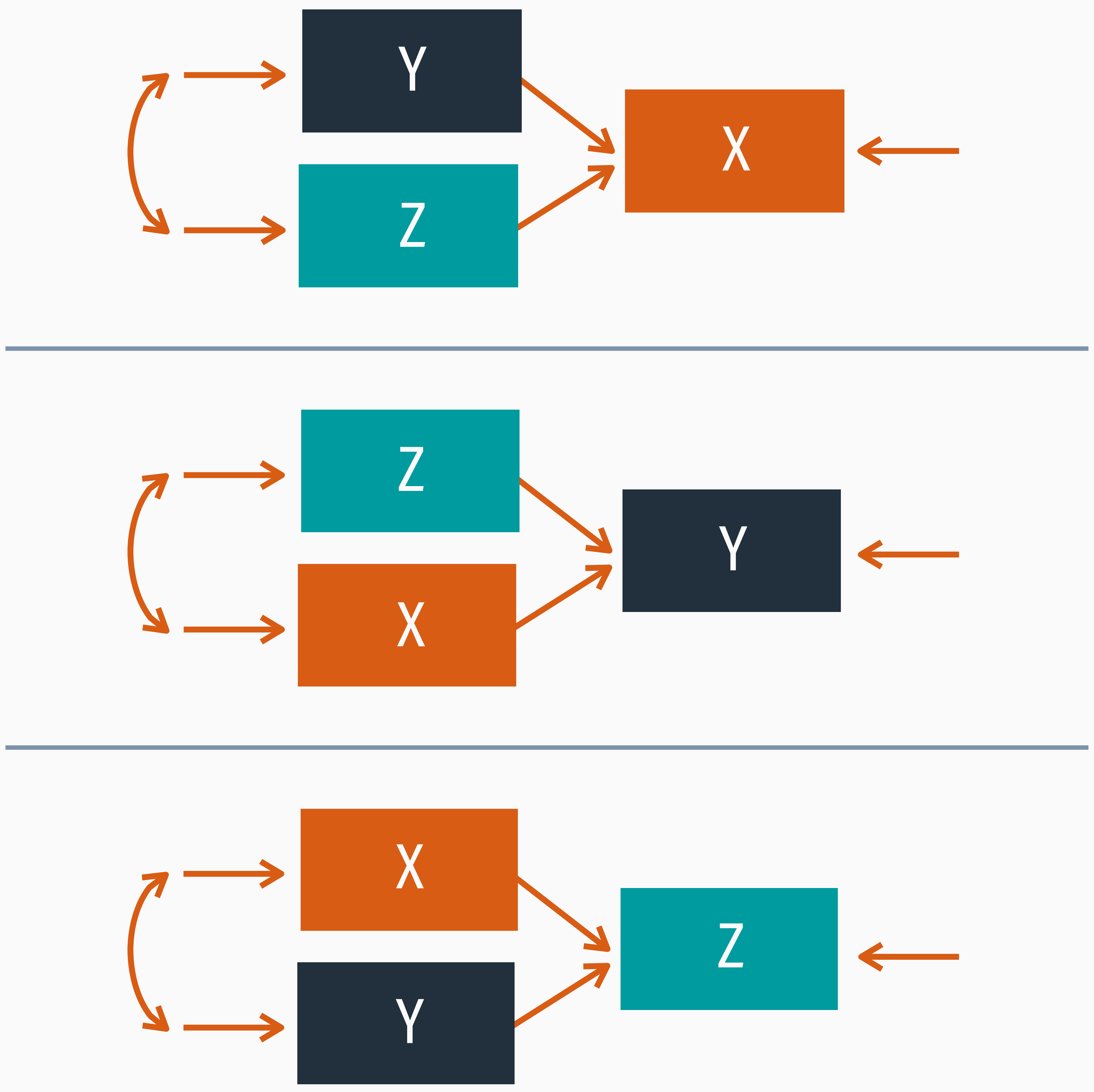
- Joint imputation invokes a multivariate distribution for the incomplete variables

- Usually a multivariate normal model with a mean vector and covariance matrix as parameters



FULLY CONDITIONAL SPECIFICATION

- FCS (also called the MICE algorithm) uses regression models to fill in data
- Each MCMC cycle uses a round-robin scheme with each variable predicted by others
- Each regression model can invoke a different metric and distribution



AGNOSTIC IMPUTATION PROS AND CONS

Pros

- Widely available in statistical software (SPSS, SAS, Stata, Mplus, R)
- Accommodates mixed metrics (normal, binary, ordinal, multicategorical)
- Can generate imputations for several purposes or analyses

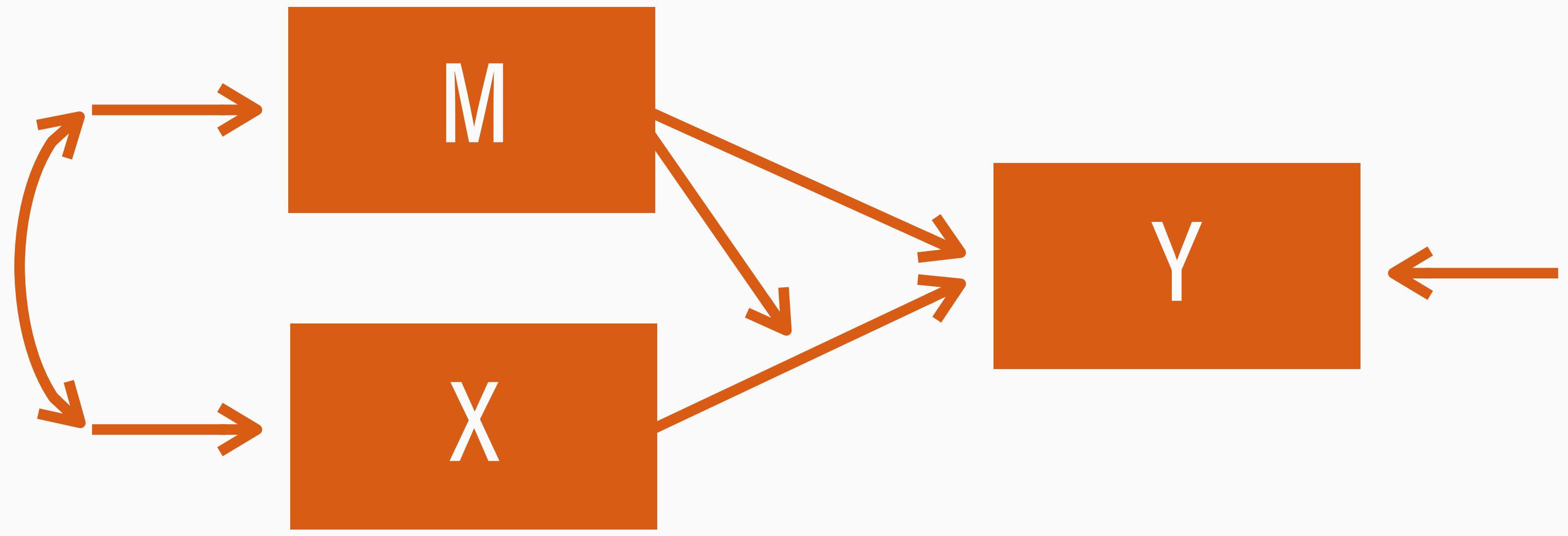
Cons

- Biased with interactions, non-linear terms, and random slope MLMs
- Capabilities vary dramatically across software packages
- Algorithms for MLMs are limited and restricted to random intercepts

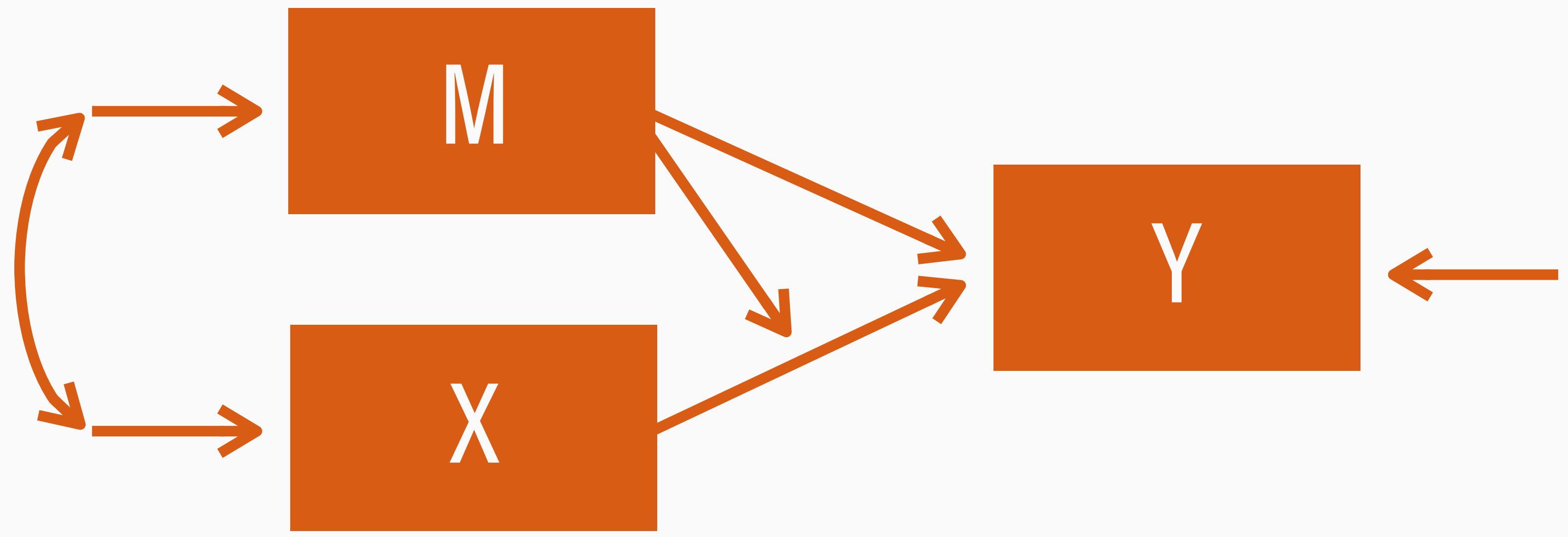
MODEL-BASED IMPUTATION

- The step 1 imputation model exactly matches the step 2 analysis model
- Imputations are tailored to one analysis, cannot be used for other purposes

Imputation model



Analysis model



MODEL-BASED IMPUTATION PROS AND CONS

Pros

- Suited for interactions, non-linear terms, and random coefficients (MLMs)
- Accommodates mixed metrics (normal, binary, ordinal, multicategorical)
- Imputation and analysis models cannot conflict or contradict each other

Cons

- Fewer simple software options (Blimp), some are difficult to use (JAGS)
- Each analysis requires a unique set of tailored imputations
- Literature on factored regression specifications is less mature

OUTLINE

SOFTWARE SUMMARY

Method	Program	Specification	Features
Maximum likelihood	Mplus	MVN limited FRS	categorical outcomes / normal predictors / robust corrections / some MLMs / latent by latent interactions
	lavaan (R)	MVN	normal variables only / robust corrections
	mdmb (R)	FRS	binary, ordinal, normal variables / manifest variable interactions
MCMC	Blimp	FRS	binary, ordinal, multicategorical, skewed, count, latent variables / MLMs / latent by latent or latent by manifest interactions
Multiple imputation	Blimp	model-based FRS agnostic FCS/MICE	model-based features are the same as MCMC FCS with normal, binary, ordinal, and multicategorical
Imputation analysis	Mplus mitml (R)	NA	multiple imputation analysis and pooling suites with test statistics

Note. MVN = multivariate normal, FRS = factored regression specification, FCS = fully conditional specification

WWW.APPLIEDMISSINGDATA.COM/BLIMP

IES Institute of Education Sciences

APPLIED MISSING DATA

home | analysis examples | blimp | blimp papers | videos | centerstat workshop | quantitude podcast

BLIMP 3.0

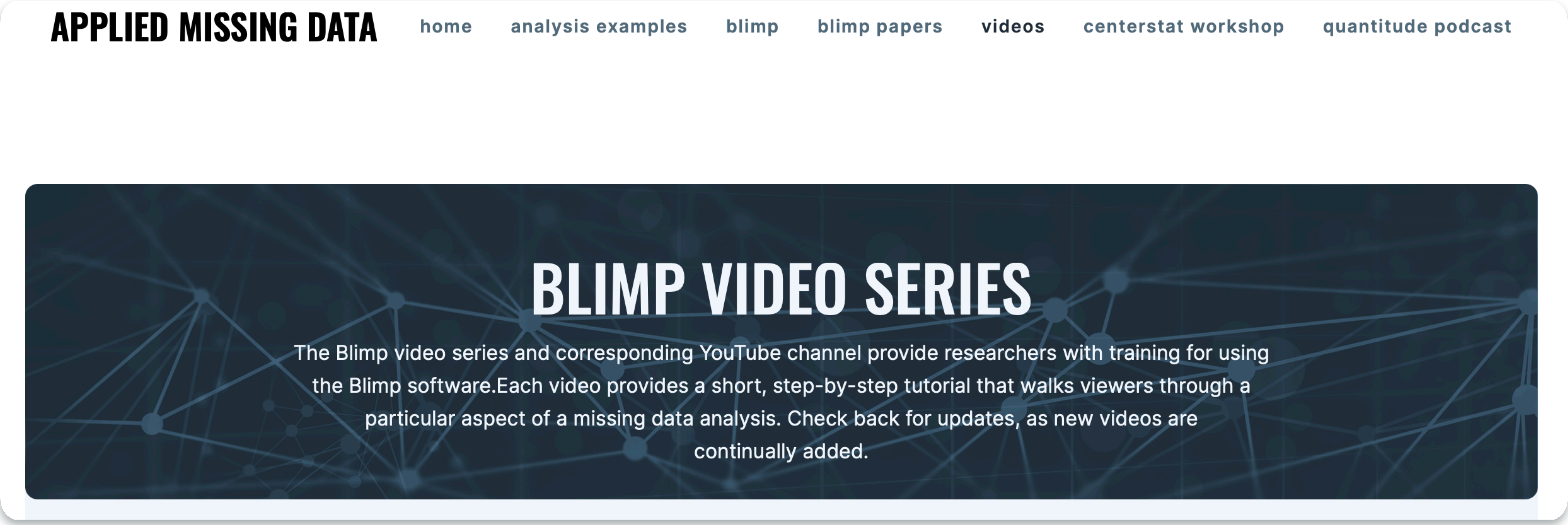
Blimp 3 offers powerful latent variable modeling and imputation for incomplete data sets with up to three levels. Blimp's unique Bayesian computational architecture allows easy specification of complex analyses that are difficult or impossible to fit in other software packages.

Download Now

User's Guide

BLIMP

WWW.APPLIEDMISSINGDATA.COM/VIDEOS



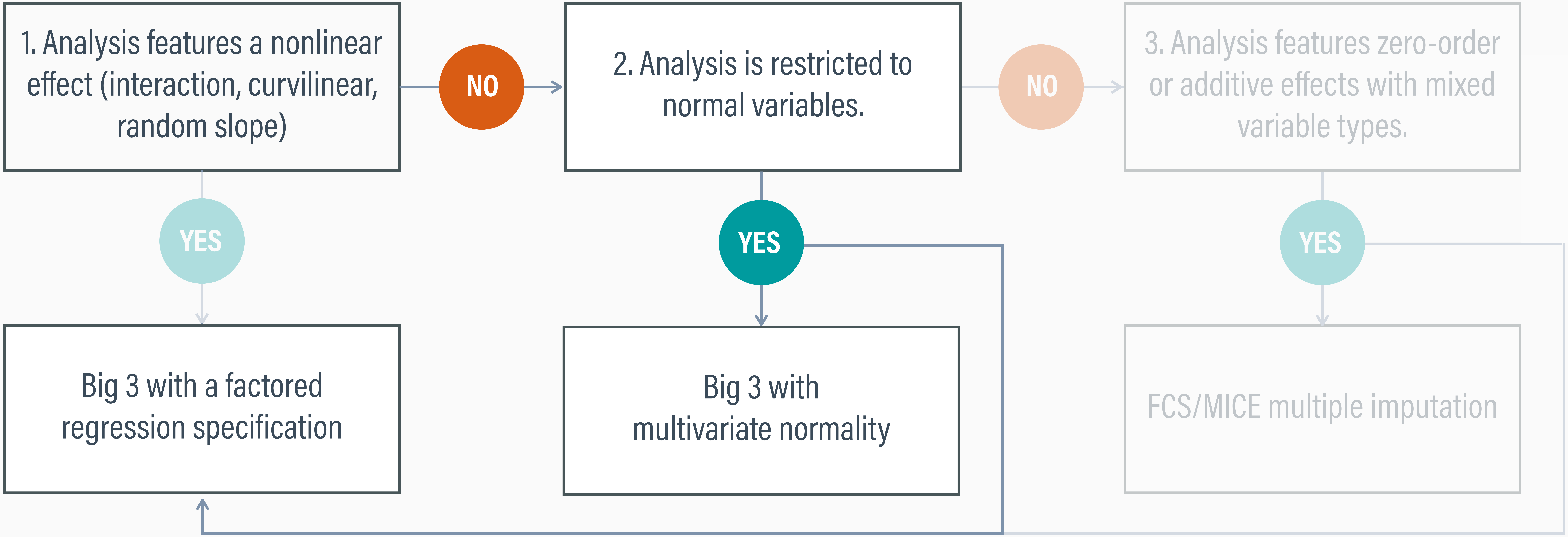
OUTLINE

ANALYSIS 1

$$read9 = \beta_0 + \beta_1(read1) + \beta_2(lrnprob1) + \varepsilon$$

Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

MISSING DATA DECISION TREE



BIG THREE COMPARISON

The Big Three are numerically equivalent!!!

	MCMC (FRS)		ML (MVN)		FCS/MICE MI	
Parameter	Est.	SD	Est.	SE	Est.	SE
Intercept	65.18	6.75	65.13	6.50	64.61	6.64
1st grade reading slope	0.51	0.05	0.51	0.05	0.51	0.05
1st grade problems slope	–0.40	0.10	–0.40	0.10	–0.40	0.09
R^2	.56	—	.57	—	.57	—

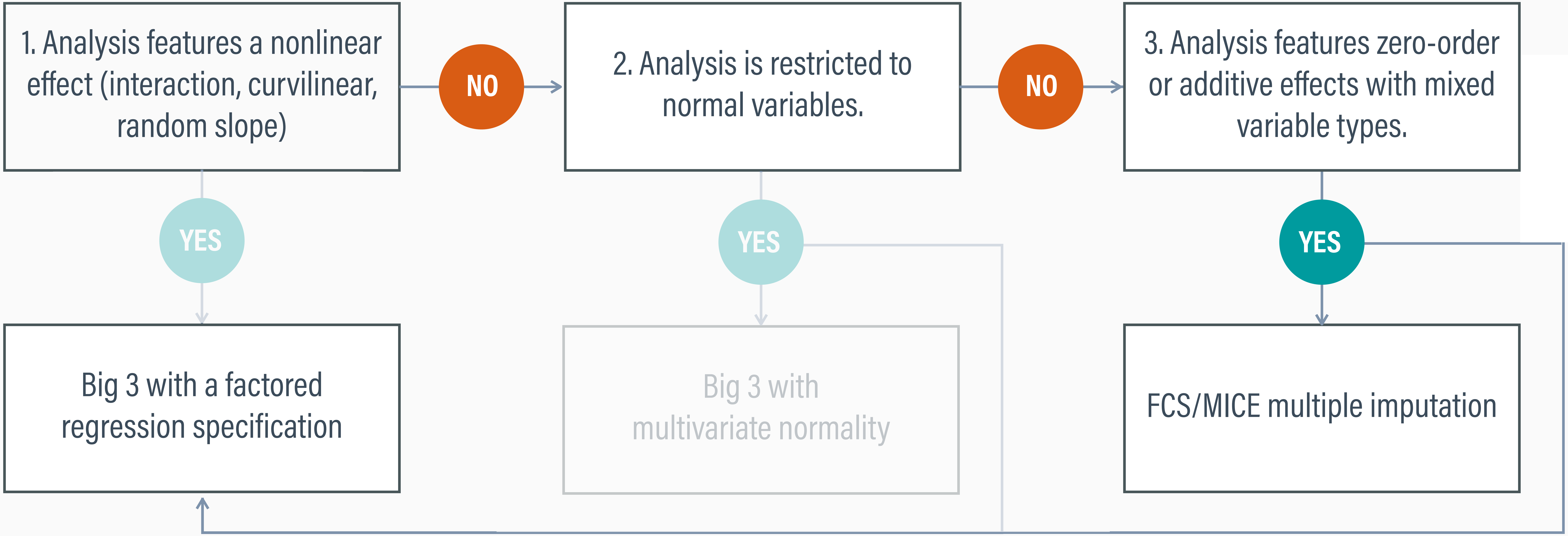
Note. FRS = factored regresson specification, MVN = multivariate normal model.

ANALYSIS 2

$$read9 = \beta_0 + \beta_1(read1) + \beta_2(lrnprob1) + \beta_3(atrisk) + \varepsilon$$

Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

MISSING DATA DECISION TREE



BIG THREE COMPARISON

The Big Three are numerically equivalent!!!

	MCMC (FRS)		ML (FRS)		FCS/MICE MI	
Parameter	Est.	SD	Est.	SE	Est.	SE
Intercept	68.47	7.25	68.43	6.96	66.87	7.12
1st grade reading slope	0.49	0.05	0.49	0.05	0.50	0.05
1st grade problems slope	−0.42	0.10	−0.42	0.09	-0.40	0.09
At risk indicator slope	−2.29	1.99	−2.27	1.93	−2.15	1.97
R^2	.57	—	.58	—	.58	—

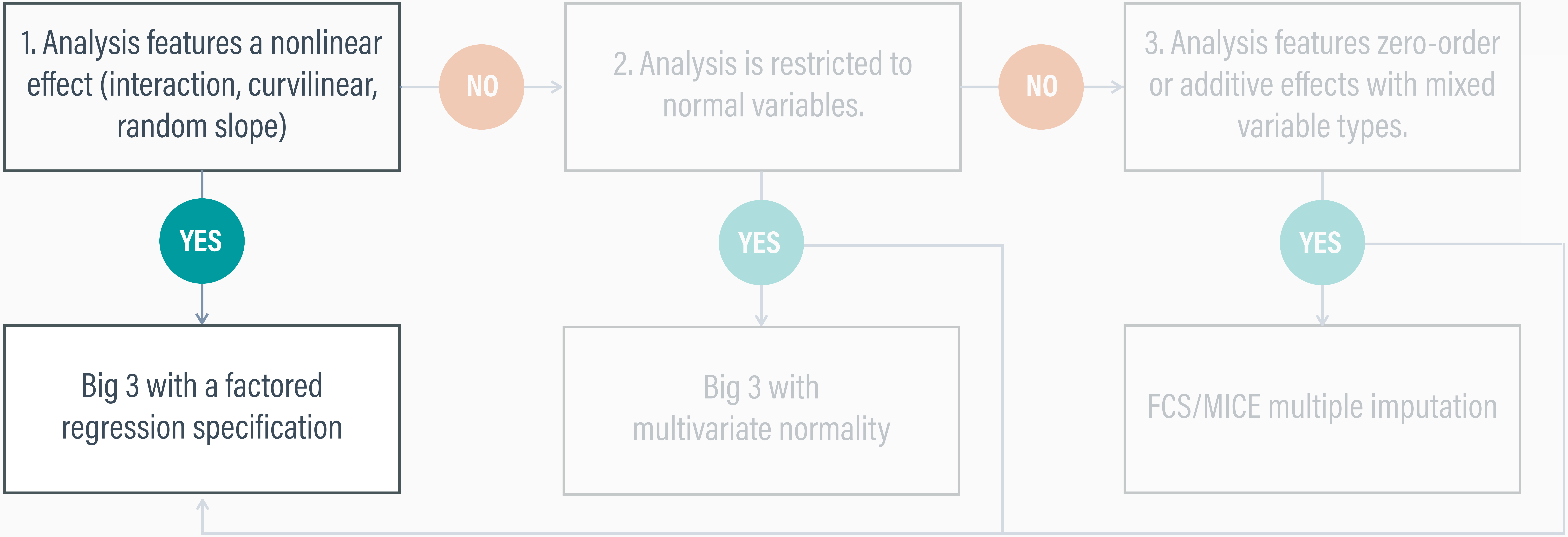
Note. FRS = factored regresson specification.

ANALYSIS 3

$$read9 = \beta_0 + \beta_1(read1) + \beta_2(lrnprob1) + \beta_3(read1)(lrnprob1) + \beta_4(atrisk) + \varepsilon$$

Variable	Definition	Missing %	Scale
atrisk	Emotional/behavioral risk code	2.2	0 = Low, 1 = Medium/high
lrnprob1	1st grade learning problems	2.2	Numeric (31 to 88)
read1	1st grade broad reading composite	6.5	Numeric (39 to 153)
read9	9th grade broad reading composite	17.4	Numeric (41 to 123)

MISSING DATA DECISION TREE



BIG THREE COMPARISON

The Big Three are numerically equivalent!!!

	MCMC (FRS)		ML (FRS)		Model-Based MI (FRS)	
Parameter	Est.	SD	Est.	SE	Est.	SE
Intercept	142.29	23.45	142.61	23.51	139.21	23.58
1st grade reading slope	–0.36	0.26	–0.36	0.26	–0.33	0.26
1st grade problems slope	–1.88	0.46	–1.89	0.46	–1.82	0.45
Reading × problems slope	0.02	0.01	0.02	0.01	0.02	0.01
At risk indicator slope	–2.05	1.91	–2.06	1.94	–1.92	1.98
R^2	.63	—	.64	—	.63	

Note. FRS = factored regresson specification.

For more information go to

WWW.APPLIEDMISSINGDATA.COM